Unit 2

Modelling variation

Introduction

A common type of experiment involves taking measurements on a sample from a population. The data introduced in Example 5 of Unit 1, for instance, are the weight changes of a sample of 83 participants in a clinical trial investigating response inhibition training. And the data in Activity 3 of that unit are the blood plasma β endorphin concentrations of a sample of 22 competitors in the Great North Run half-marathon. As another example of a sample, the data in Table 1 below are the lengths (in cm) of a sample of 100 leaves from an ornamental bush.

Table 1 Leaf lengths (cm)

1.6	1.9	2.2	2.1	2.2	1.0	0.8	0.6	1.1	2.2
1.3	1.0	1.1	0.8	1.4	2.2	2.1	1.3	1.0	1.3
1.1	2.1	1.1	1.1	1.0	0.9	1.3	2.3	1.3	1.0
1.0	1.3	1.3	1.5	2.4	1.0	1.0	1.3	1.1	1.3
1.3	0.9	1.0	1.4	2.3	0.9	1.4	1.3	1.2	1.5
2.6	2.7	1.6	1.0	0.7	1.7	0.8	1.3	1.4	1.3
1.5	0.6	0.5	0.4	2.7	1.6	1.1	0.9	1.3	0.5
1.6	1.2	1.1	0.9	1.2	1.2	1.3	1.4	1.4	0.5
0.4	0.5	0.6	0.5	0.5	1.5	0.5	0.5	0.4	2.5
1.6	1.5	2.0	1.4	1.2	1.6	1.4	1.6	0.3	0.3

The measurements in each of these examples vary: weight change varies from trial participant to trial participant; blood plasma β endorphin concentration varies from runner to runner; and leaf length varies from leaf to leaf. Furthermore, if we decided to obtain another measurement, we could not predict exactly what that measurement would be: we could not say what the weight loss of another participant in response inhibition training would be for sure, nor what the blood plasma β endorphin concentration of another runner would be, nor how long another leaf from the ornamental bush would be. Because their measurements vary, weight change, blood plasma β endorphin concentration and leaf length are **random variables**.

A random variable may take any value from a set of possible values, although some values may be more likely than others to occur. Consider, for instance, the leaf lengths of Table 1. From the frequency histogram in Figure 1 (overleaf), it is clear that not many of the leaves in the sample were more than 2.5 cm long, whereas quite a large proportion of the leaves were between 0.5 cm and 2.0 cm long. So if another leaf were to be taken at random from the same bush and measured, we might feel it was more likely to be between 0.5 cm and 2.0 cm long than it was to be longer than 2.5 cm.

As usual, in this histogram, leaves whose recorded lengths were exactly 0.5 cm (say) have been allocated to the class interval 0.5–1.0 cm, leaf lengths recorded as 1.0 cm to the class interval 1.0–1.5 cm, and so on.

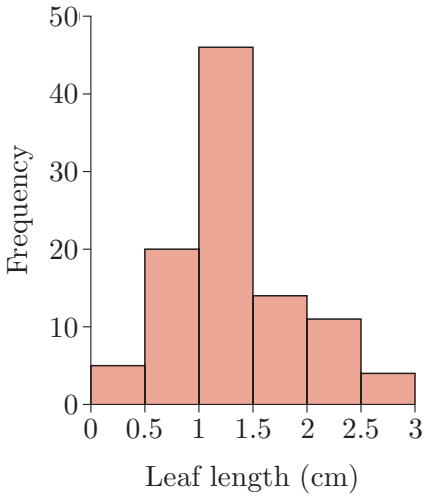


Figure 1 Histogram of leaf lengths

A concept which is essential for modelling this variability is that of *probability*: basically, a probability is a number which measures how likely an event is to occur. In Section 1, probability is defined, some notation is introduced and some basic properties of probabilities are discussed briefly. Section 2 is concerned with general features of and ideas about modelling random variables. In particular, the distinction between *discrete* and *continuous* random variables is emphasised. The general notion of a *probability distribution* is introduced in Subsection 2.2. It turns out that probability distributions are characterised a little differently for discrete and continuous random variables, the former via the notion of a *probability mass function*, the latter through a *probability density function*. Some of the ideas involved are illustrated using computer simulations.

In this module, you will need to integrate only powers, polynomials and exponential functions; in this unit, only powers and polynomials.

In the continuous case, the use of integration to calculate probabilities and to check the validity of a probability density function is described in Section 3, after first revising the results on integration that are required here. Finally, in Section 4, a further important function is introduced, the *cumulative distribution function*. This function allows you to calculate numerous probabilities without huge effort. It is defined for both discrete and continuous random variables, in the latter case once again making use of the integration techniques you used earlier in the unit.

1 What is probability?

The idea of using a number – a *probability* – to measure how likely a chance event is to occur emerged towards the end of the seventeenth century. On the continent of Europe, the desire of some gamblers to analyse various games of chance, particularly those involving dice, led to the development of a theory of probability. In England, at about the same time, a different approach to analysing chance events was adopted; this was based on the

collection of data. In this section, these two approaches to measuring how likely a chance event is to occur are introduced.

In the first approach, probabilities can be deduced from assumptions about the situation, such as symmetry. For instance, if a six-sided die is fair, or *unbiased*, then each of its six sides is equally likely to be the one that is uppermost when it is rolled. Thus the probability that it will land with a particular face uppermost is 1/6; for example, the probability that a four will be rolled is 1/6.

Actually, there's a third important approach to probability which is a subjective one for use with non-repeatable events; this will not be pursued in this module.

Gambling with dice in medieval times

Example 1 *Coin tossing*

Suppose that a coin is tossed a large number of times and we are interested in the event 'the coin lands with its heads side uppermost' or 'the coin lands heads' or 'heads' for short. Since there is no reason to believe that either heads or tails is more likely to occur than the other, you would expect the coin to land 'heads' for approximately half of the tosses (and 'tails' for approximately half of the tosses). Thus the probability of a head is 1/2.

In Example 1, we made the usual assumption that the coin cannot land on its edge. The following is an aside that shows that this doesn't always apply!

> Has any university department ever opened its account with such a statistically significant event as that which launched the Warwick Statistics Department? On Tuesday 9th October 1972, in the first serious lecture given to a group of 45 second-year mathematicians, entitled Possibilities & Probabilities, the founding professor tossed a 2p coin high in the air. The coin descended to the vinyl floor of lecture theatre L5, spun as a perfect sphere, and, in full view, slowly came to rest on its edge! Stunned silence turned into massive applause. No further publicity was necessary – truly the Statistics Department had arrived in style!
>
> (Source: Harrison, J. 'A Brief History of the Early Years of the Statistics Department', https://www2.warwick.ac.uk/fac/sci/maths/general/institute)

Activity 1 *Roulette wheels*

A European roulette wheel has 37 equal-sized compartments numbered from 0 to 36. Suppose that the wheel is fair – that is, there is no reason to suppose that the ball is more or less likely to come to rest in one compartment than in any other.

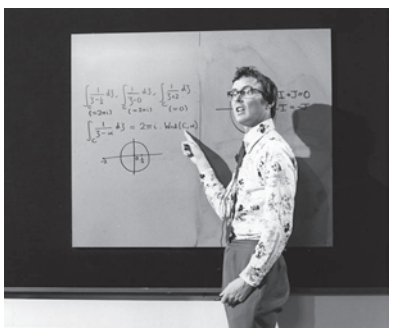
(a) What is the probability that the ball will come to rest in the compartment numbered 19?

(b) What is the probability that the ball will come to rest in an odd-numbered compartment?

The other main approach to obtaining a probability is to use data. The basic tenet of this approach is summed up in the following box.

Probability is equivalent to Proportion.

This idea will be used in two ways in this section. In the first, suppose we randomly pick an individual (person or item) from a given finite set of individuals, the latter comprising our dataset. Then the probability that the randomly chosen individual has a particular characteristic is equal to the proportion of individuals in the set that have that characteristic. This is illustrated in Example 2.



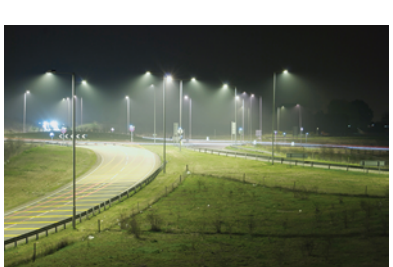
Example 2 *Genders of academics*

At the end of 2015, the Department of Mathematics and Statistics at The Open University contained 43 'permanent' academics, of whom 18 were women and 25 were men. The proportion of female academics in this department was therefore $18/43 \simeq 0.42$. It is also the case that if an academic from this department were to be selected at random to appear on a television programme, say, the probability that the academic selected was female would be $18/43 \simeq 0.42$.

Activity 2 *Colour blindness*

In a class of 25 pupils, two are colour blind. What is the probability that a pupil picked at random from the class is colour blind?

In Example 2 and Activity 2, the datasets in question – a department of 43 academics, a class of 25 schoolchildren – were treated as if they were *populations*. We were then able to answer questions about probabilities associated with individuals chosen randomly out of those small populations. But, as discussed in Unit 1 and the Introduction, more usually datasets are themselves *samples* of individuals that have been randomly selected from some much larger underlying population. Interest then is in inferring something about that population based on information provided by the sample. In particular, this is the second way in which we think of probabilities as proportions: the proportion of individuals in a sample with a given characteristic is an *estimate* of the probability that an individual in a larger population has the characteristic. This is illustrated in Example 3.



Example 3 *Faulty street lamps*

Suppose that a local council suspects that the latest consignment of LED lights for its street lamps is of poor quality, with an unacceptably large proportion of them being faulty and not working. To investigate this, a council official decides to examine a sample of lights from this consignment.

The official enters the warehouse where the rest of the consignment is stored and randomly chooses lights: that is, he chooses lights in such a way that no light is more likely or less likely to be chosen than any other light. He stops when he obtains 100 lights.

Among the 100 lights chosen, five do not work. That is, the proportion $5/100 = 0.05$ of these 100 lights do not work. Had the official only been interested in these particular 100 lights, he could then say that the probability that a street light randomly chosen from these 100 street lights is faulty is 0.05.

But the official is really interested not in these 100 lights as such, but in what they tell him about the number of faulty lights in the whole consignment. Then the proportion of lights observed to be faulty in this sample – 0.05 – is an estimate of the proportion of faulty lights in the consignment.

This example is typical of a particular problem of quality control: the estimation of 'percentage defectives' in a batch of supplied items.

In fact, the value 0.05 is also an estimate of the probability that any such LED street light, not just from this particular consignment, will be faulty. It will be a bad estimate of this probability if this particular consignment is not typical of such street lights.

Activity 3 *Credit card debt*

A random sample of 2000 adults living in the UK were surveyed about their financial situation. Of these, 474 reported that they have outstanding credit card debts. Estimate the probability that an adult living in the UK has outstanding credit card debts.

Activity 4 *Helping behaviour*

In an experiment conducted some years ago to explore the issue of whether people are generally more helpful to females than to males, eight students approached people and asked if they could change a 5p coin. Altogether 100 people were approached by the male students and 105 by the female students. The results of the experiment are displayed in Table 2.

The amount 5p may seem very small and prompt the question why change might be required for such a small sum. At the time the experiment was carried out, a local telephone call could be made from a public telephone box for as little as 2p.

Table 2 Helping behaviour

Sex of student	Help given	Help not given
Male	71	29
Female	89	16

(Source: Sissons, M. (1981) 'Race, sex and helpful behaviour', *British Journal of Social Psychology*, vol. 20, no. 4, pp. 285–92)

(a) Use these data to estimate the probability that a male will be given help under these circumstances.

(b) What would you estimate the probability to be for a female?

(c) Do the results of the experiment support the notion that people are more helpful to females?

In Activity 4, you were asked to estimate two probabilities. One of the main themes of this module is estimation – obtaining estimates and assessing how reliable these estimates are. In this activity, you were also asked to comment on the meaning of the results of the experiment. Formal ways of quantifying the extent to which experimental results support a claim, or hypothesis, are also discussed later in the module.

1.1 Formalising the notion of probability

We will now define probabilities more formally. In general, suppose that an event E (say) may or may not occur in an experiment – that is, the outcome of the experiment is uncertain (it is not possible to say beforehand what will happen) – and suppose that the experiment can be repeated (at least in principle) as often as we like. For instance, the event might be obtaining a four when a die is rolled, or obtaining a head when a coin is tossed. If the experiment is repeated many times, then the *number* of times the event E occurs is the **sample frequency** of the event E, and the *proportion* of times that E occurs is the **sample relative frequency** of the event E.

If the experiment is repeated *an enormous number of times*, then we can think of the relative frequency as the **probability** that the event E occurs. More formally, the probability that the event E occurs is the proportion towards which the sample relative frequency is tending as we increase the number of times the experiment is repeated. This probability is denoted by $P(E)$; this is usually read as 'the probability of E' or simply as 'P of E'.

There is a 'settling down' notion here: as an experiment or situation is repeated more and more times, the proportion of the time that a particular event occurs 'settles down' to a particular value, which is the probability of that event occurring. This notion is explored in Screencast 2.1.

Screencast 2.1 Proportions settling down to probabilities

Activity 5 *Values of probabilities*

As the probability of E is a proportion, what do you think can be said about the possible values that $P(E)$ can take?

In addition to the set of possible values for $P(E)$, two other properties of $P(E)$ are immediate. If an event is impossible, then it never happens, so its probability is 0; and if an event is certain, then it always happens, so its probability is 1. We can summarise these results as follows.

Properties of probabilities

- For any event E, $0 \leq P(E) \leq 1$.
- If an event E is impossible, then $P(E) = 0$.
- If an event E is certain to happen, then $P(E) = 1$.

You can use the first property as a 'common sense' check in probability calculations: if you obtain a value for a probability outside the interval 0 to 1, then you will know that you have made a mistake in your calculations.

A further property of probabilities, that will be used in Section 4, arises directly from the above. Because an event either occurs or does not occur, it must be the case that $P(E \text{ occurs}) + P(E \text{ does not occur}) = 1$. Rearranging this equation gives the following rule; the event 'E does not occur' is called the *complementary event* to E.

The text is about *complementary* events; this ticket is for a *complimentary* event!

Probability rule for complementary events

For any event E,

$$P(E \text{ does not occur}) = 1 - P(E \text{ occurs}).$$

One further property of probabilities, which will be used in later units of the module, concerns the probability of more than one event. Suppose that we now have two events E_1 and E_2, where the probability that E_1 occurs does not affect the probability that E_2 occurs, and vice versa. In this case, the two events are said to be **independent**.

For two independent events E_1 and E_2, the probability that both events occur is

$$P(E_1 \text{ and } E_2) = P(E_1) \times P(E_2).$$

Note that this is true only if E_1 and E_2 are independent.

Example 4 *Two rolls of a die*

A fair six-sided die is rolled twice. Let E_1 be the event that a six lands uppermost, and let E_2 be the event that any number other than six lands uppermost.

Because the die is fair,

$$P(E_1) = \frac{1}{6}.$$

Event E_2 is the complementary event of E_1, so

$$P(E_2) = 1 - P(E_1) = 1 - \frac{1}{6} = \frac{5}{6}.$$

The outcomes of the two die rolls are independent since the outcome of any die roll is unaffected by the outcome of any other die roll. So the probability of rolling a six on the first roll, and any number other than six on the second roll, is

$$P(E_1 \text{ and } E_2) = P(E_1) \times P(E_2) = \frac{1}{6} \times \frac{5}{6} = \frac{5}{36}.$$

This can be extended to a general result for the probability of r independent events $E_1, E_2, \ldots, E_r$.

Probability rule for multiple independent events

For any r independent events $E_1, E_2, \ldots, E_r$, the probability that all the events occur is

$$P(E_1 \text{ and } E_2 \text{ and } \ldots \text{ and } E_r) = P(E_1) \times P(E_2) \times \cdots \times P(E_r).$$

Exercises on Section 1

Exercise 1 *Dice and symmetry*

(a) A tetrahedron is a regular four-sided solid, with each face an equilateral triangle. A tetrahedral die has faces labelled 1, 2, 3 and 4. The die is rolled. Assuming that the die is unbiased, what is the probability that it will come to rest on the face labelled 3?

(b) An octahedron is a regular eight-sided solid, with each face an equilateral triangle. An octahedral die has faces labelled $1, 2, 3, \ldots, 8$. The die is rolled. Assuming that the die is unbiased, what is the probability that it will come to rest on a face labelled either 3 or 6?

(c) A tetrahedral die and an octahedral die are rolled. What is the probability that the tetrahedral die will come to rest on the face labelled 3, and the octahedral die will come to rest on a face labelled either 3 or 6?

Exercise 2 *Tiger beetles*

Cicindela fulgida: bright red or not bright red?

The colour patterns of 671 tiger beetles of the genus *Cicindela fulgida* were classified as either bright red or not bright red. (Source: Sokal, R.R. and Rohlf, F.J. (2012) *Biometry*, 4th edn, New York, W.H. Freeman, p. 753.) Of the beetles found in the spring, 302 were bright red and 202 were not bright red. Of those found in the summer, 72 were bright red and 95 were not bright red.

(a) Use these data to estimate the probability that a tiger beetle found in the spring will be bright red.

(b) What is your estimate of the probability that a tiger beetle found in the summer will not be bright red?

(c) What is your estimate of the probability that a tiger beetle found in the summer will be bright red? Find the value of this estimate in two ways: directly and by using the probability rule for complementary events.

2 Modelling random variables

In the Introduction it was noted that a random variable may take any value from a set of possible values. When that set contains only a discrete set of values (such as $0, 1, 2, \ldots$), we have a *discrete* random variable, while the random variable is *continuous* if it can take any value within a continuous range of values (such as $(0, \infty)$). The distinction between discrete and continuous random variables, already made in Section 1 of Unit 1, is important and is discussed in Subsection 2.1. It will, for example, affect how we model a random variable and how we describe the probabilities associated with a random variable.

The word 'discrete' differs from 'discreet' meaning 'circumspect' and 'unobtrusive'.

Probabilities, as discussed in Section 1, are central to models associated with random variables. The probabilities are given by a *probability mass function* for a discrete random variable and through a *probability density function* for a continuous random variable. These functions are introduced in Subsections 2.2 and 2.3, respectively.

2.1 Discrete and continuous random variables

The set of possible values that a random variable can take is called the **range** of the random variable. (In other texts, the range might be called the 'sample space' or the 'support' of the distribution.) The following are examples of **discrete random variables** as each has a range that is a discrete set of values.

Notice that this usage of the term 'range' is not quite the same as its use in a sampling context: the range of a sample is the difference between the maximum sample value and the minimum sample value. Here, it is just all possible values.

Example 5 *Defective hinges*

A manufacturer produces hinges in batches of 1000. If X denotes the number of defective hinges in a batch, then X is a discrete random variable whose range is $\{0, 1, 2, \ldots, 1000\}$.

Example 6 *Waiting to join in*

In some board games, a player cannot join in until he or she has obtained a six on the roll of a die. The number of rolls necessary to obtain a six is a random variable N (say). A player may obtain a six for the first time on the first roll, or the second, or the third, or the fourth, and so on. Or it may require a very large number of rolls to obtain a six: extremely high values are unlikely, but they are not impossible. The range of N is $\{1, 2, 3, 4, \ldots\}$; it is a discrete set that contains an infinite number of values.

Activity 6 *The score on a die*

When a six-sided die is rolled, the value on the face showing uppermost is a discrete random variable. What is the range of this random variable?

Often the value taken by a discrete random variable results from a 'count', as in Examples 5 and 6. As you have also just seen, the range of a discrete random variable can be finite, as in Example 5 and Activity 6, or infinite, as in Example 6.

Yet other discrete random variables arise even when the outcome of a study is not immediately given in numerical form, but is 'coded' to do so. An example of a particular but very important sort is that of a 'binary' random variable, as given in Example 7.

Example 7 *Cured or not cured?*

It is convenient and usual for random variables to take numerical values, so even when the outcome of an experiment is non-numerical, we typically code outcomes as numbers. So, for example, if the result of a medical treatment is either 'cured' or 'not cured', we might define a random variable, X say, that takes the value 0 if a patient is cured and 1 if the patient is not cured. Thus X is a discrete random variable whose range is $\{0, 1\}$.

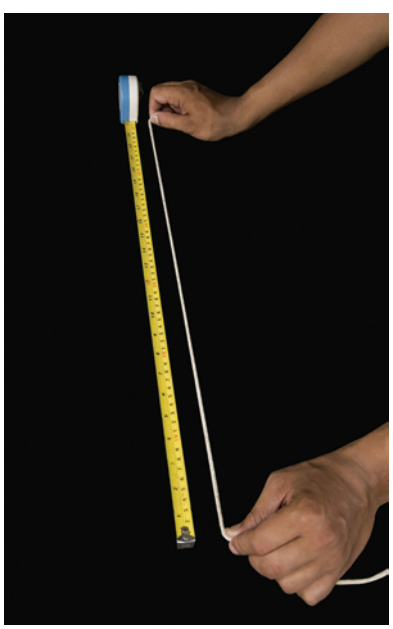
The value of a **continuous random variable**, on the other hand, is typically obtained by a direct 'measurement', as in Example 8.

Example 8 *Leaf lengths*

The leaf lengths of Table 1 are repeated in Table 3.

Table 3 Leaf lengths (cm)

1.6	1.9	2.2	2.1	2.2	1.0	0.8	0.6	1.1	2.2
1.3	1.0	1.1	0.8	1.4	2.2	2.1	1.3	1.0	1.3
1.1	2.1	1.1	1.1	1.0	0.9	1.3	2.3	1.3	1.0
1.0	1.3	1.3	1.5	2.4	1.0	1.0	1.3	1.1	1.3
1.3	0.9	1.0	1.4	2.3	0.9	1.4	1.3	1.2	1.5
2.6	2.7	1.6	1.0	0.7	1.7	0.8	1.3	1.4	1.3
1.5	0.6	0.5	0.4	2.7	1.6	1.1	0.9	1.3	0.5
1.6	1.2	1.1	0.9	1.2	1.2	1.3	1.4	1.4	0.5
0.4	0.5	0.6	0.5	0.5	1.5	0.5	0.5	0.4	2.5
1.6	1.5	2.0	1.4	1.2	1.6	1.4	1.6	0.3	0.3



How long is a piece of string? To what accuracy?

Notice that each length is given in centimetres and each measurement is recorded correct to one decimal place, that is, it is recorded to a whole number of millimetres. However, leaves do not come in exact millimetre lengths. Although the lengths are recorded to the nearest millimetre, even if we were able to measure leaf lengths to an amazing degree of accuracy the actual length of a leaf will almost certainly not be exactly equal to the recorded value – although the difference may be tiny (maybe only $0.00\ldots01$ cm!). The range of leaf lengths constitutes a continuum of values between some minimum and maximum values.

Other examples of continuous random variables are the age of an elephant, the weight of a bag of beans, the height of a building, a person's systolic blood pressure, and so forth. These too will typically be given in rounded, or 'discretised', form: an elephant's age might be recorded as 35.5 years, the weight of a bag of beans might be recorded as 0.47 kg, and so on. This raises the question, why not treat every set of data as discrete? The answer is that it turns out to be simpler and more informative when modelling data to treat them as continuous if they arise from measurements on a continuous scale, and to use models for discrete data only when the range of the data is really discrete.

The following activity will give you some further practice at distinguishing between discrete and continuous random variables.

Activity 7 *Discrete or continuous?*

(a) Table 4 gives the lengths (in mm) of the jawbones of 23 kangaroos of the *Macropus giganteus* species.

Table 4 Jawbone lengths (mm)

108.6	115.8	113.1	109.0	117.5	90.1	108.4	114.9
106.9	124.0	134.5	117.9	130.9	144.3	133.9	136.1
137.7	125.3	129.3	153.9	153.0	152.6	154.7	

(Source: Andrews, D.F. and Herzberg, A.M. (1985) *Data*, New York, Springer-Verlag, p. 311)

Would you choose to model jawbone length as a discrete or a continuous random variable?

(b) Table 5 gives the number of yeast cells found in each of 400 very small squares on a microscope slide when a liquid was spread over it. The first row gives the number x of yeast cells observed in a square, and the second row gives the number of squares containing x cells for each value of x. For instance, 213 of the 400 squares did not contain any yeast cells. No square contained more than 5 cells.

Table 5 Yeast cells on a microscope slide

Cells in a square, x	0	1	2	3	4	5
Frequency	213	128	37	18	3	1

(Source: 'Student' (1906) 'On the error of counting with a haemocytometer', *Biometrika*, vol. 5, no. 3, pp. 351–60)

The number of yeast cells per square is a random variable taking (in this experiment) observed values between 0 and 5. Would you model the variation using a discrete or a continuous probability model?

(c) On Thursday 23 June 2016, a referendum was held in the UK on the issue of whether or not it should remain a member of the European Union (EU). Ignoring rejected ballots, define Y to be 1 if an individual voted to remain in the EU, and define Y to be 0 if the individual voted to leave the EU. The results of the vote are shown in Table 6.

Table 6 Votes to remain and to leave the EU

Outcome	0	1
Number of votes	17 410 742	16 141 241

Is Y a discrete or a continuous random variable?

(d) Table 7 contains the times, in minutes, at which the insulation failed for 12 electrical components of a particular type subject to increasing voltages.

Table 7 Failure times of electrical insulation (minutes)

219.3	79.4	86.0	150.2	21.7	18.5
121.9	40.5	147.1	35.1	42.3	48.7

(Source: Lawless, J.F. (2003) *Statistical Models and Methods for Lifetime Data*, 2nd edn, Hoboken, NJ, Wiley-Interscience, p. 208)

What sort of model would you adopt for the variation in failure times?

(e) The analysis of spontaneous (fossil) fission tracks can be used as a dating method on geological timescales. To quote from the source of the data: 'Fission tracks are trails of damage in the crystal structure of a mineral, caused by the fissioning of uranium atoms.' Table 8 shows the number of spontaneous fission tracks in each of 30 grains of apatite found in Mahe granite, Seychelles.

Table 8 Numbers of fission tracks

0	2	18	2	10	3	4	20	52	2
1	6	256	52	3	10	2	7	1	14
15	14	8	22	16	34	14	6	13	127

(Source: Gleadow, A. in Galbraith, R.F. (2005) *Statistics for Fission Track Analysis*, Boca Raton, LA, Chapman & Hall/CRC, p. 34)

The number of spontaneous fission tracks per grain of apatite is a random variable. Is the random variable discrete or continuous?

2.2 Probability distributions and probability mass functions

A **probability distribution** links each possible value of a random variable with its probability of occurrence.

Example 9 *Probability distribution for cured or not cured*

In Example 7, we defined the (binary) discrete random variable X to be 0 if a patient is cured and 1 if the patient is not cured. With this coding,

$$P(\text{cured}) = P(X = 0) \quad \text{and} \quad P(\text{not cured}) = P(X = 1).$$

So if the treatment cures three-quarters of patients, say, then the probability distribution of X is

$$P(X = 0) = 3/4, \quad P(X = 1) = 1/4.$$

In general, if we represent the outcome of a study by a random variable, then we can express the probability distribution for the range of possible outcomes using a mathematical function. For *discrete* random variables this function is called the **probability mass function**. It is normally denoted by the lower-case letter p, so for each x in the range of the random variable X, we have

$$p(x) = P(X = x).$$

Note the difference between the use of the lower-case letter p and the upper-case letter P in a probability context. The notation $P(.)$ is used exclusively to represent the phrase 'the probability that' with reference to an event; you should not use an upper-case letter P for anything else. On the other hand, the lower-case letter p is the name of a probability function; and $p(x)$ is read simply 'p of x'. (A probability mass function is always denoted by a lower-case letter – usually p, although other letters are sometimes used.)

Notice also the convention that an upper-case letter (X, for example) is used for the label of a random variable, while the corresponding lower-case letter (x) is used as representative of the possible values the random variable might take.

The following examples show common ways of representing a probability mass function.

Example 10 *Probability mass function for cured or not cured*

As in Examples 7 and 9, let $X = 0$ and $X = 1$ denote 'cured' and 'not cured', respectively. From Example 9, $P(X = 0) = 3/4$ and $P(X = 1) = 1/4$. So the probability mass function associated with X can be written as

$$p(x) = \begin{cases} 3/4 & x = 0 \\ 1/4 & x = 1. \end{cases}$$

This probability mass function can be depicted in a simple graph; see Figure 2.

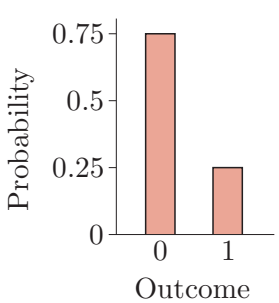


Figure 2 The probability mass function for a random variable representing 'cured' and 'not cured'

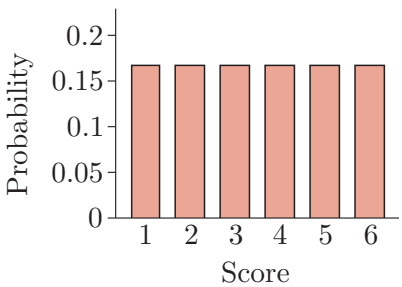


Figure 3 The probability mass function for an unbiased die

Example 11 *The score on an unbiased die*

Suppose that the random variable Y represents the score obtained when an unbiased six-sided die is rolled. The range of Y is $\{1, 2, 3, 4, 5, 6\}$, and each value has probability $1/6$ of occurring. The probability mass function of Y may be written as

$$p(y) = 1/6, \quad y = 1, 2, 3, 4, 5, 6.$$

Notice that the upper-case letter Y has been used for the name of the random variable and the corresponding lower-case letter y for possible observed values of the random variable. The probability mass function may also be shown in a diagram, as in Figure 3.

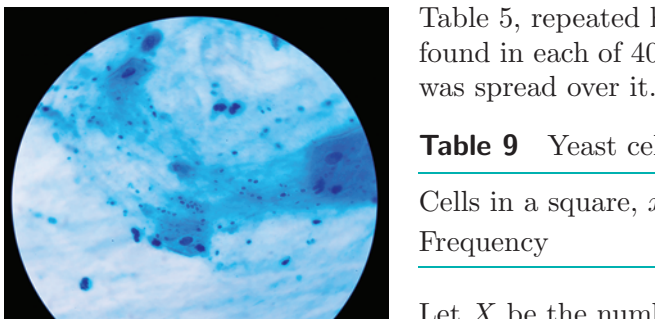
Example 12 *Yeast cells*

Table 5, repeated here for convenience, gave the number of yeast cells found in each of 400 very small squares on a microscope slide when a liquid was spread over it.

Table 9 Yeast cells on a microscope slide

Cells in a square, x	0	1	2	3	4	5
Frequency	213	128	37	18	3	1

Let X be the number of yeast cells contained in one of the squares *picked at random* from the whole set of 400 small squares. The probability that the randomly chosen square contains no yeast cells would be $p(0) = 213/400 = 0.5325$, the probability that it contains one yeast cell would be $p(1) = 128/400 = 0.32$, and so on. A table could be used to show the probability mass function of X, as could a figure. These are shown in Table 10 and Figure 4.



Yeast cells

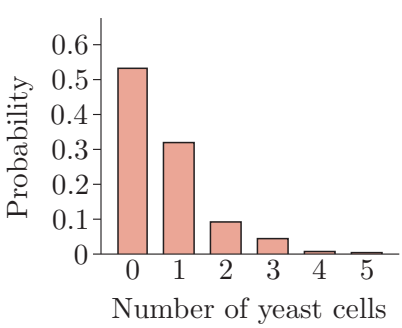


Figure 4 The probability mass function given in Table 10

Table 10 Probability mass function for yeast cells

x	0	1	2	3	4	5
$p(x)$	0.5325	0.32	0.0925	0.045	0.0075	0.0025

Activity 8 *Probability mass functions*

(a) In Example 3, the following scenario was considered. An official from a local council checked a sample from a consignment of LED street lights to see whether or not they were faulty. He found that 5% of the sample of street lights were faulty. Suppose that, unbeknown to the official, 4% of the street lights in the entire consignment were faulty. Write down an appropriate random variable reflecting the faultiness or otherwise of a light selected at random from the consignment, and give its probability mass function.

(b) Suppose that each face of a six-sided die is equally likely to be uppermost when the die is rolled but (unlike an ordinary die) two of its faces show a '5' and its other faces show 1, 3, 4 or 6. If X is the uppermost number after rolling the die, give the probability mass function of X.

The next box summarises the definition of the probability mass function, together with some associated terminology.

The probability mass function

The probability function for a *discrete* random variable is usually called the **probability mass function** (or simply the **mass function**) of the random variable. This is often abbreviated to **p.m.f.** For a discrete random variable X, the probability mass function gives the probability distribution of X:

$$p(x) = P(X = x).$$

The p.m.f. is defined for all values x in the range of X.

In Subsection 1.1, we explored the notion of the relative frequency of an event observed in a sample settling down and becoming a probability as an experiment is repeated an enormous number of times. This idea applies to the sample relative frequencies, and hence probabilities, of each event of the form '$X = x$'. In this way, for all x in the range of X, the whole set of sample relative frequencies of occurrences of the values of x settles down towards the whole set of probabilities that $X = x$. And this set of probabilities is what we defined to be the probability mass function above. The idea of a whole sample settling down towards a probability model as the sample size increases is explored for a discrete distribution in Chapter 5 of Computer Book A.

Refer to Chapter 5 of Computer Book A for the next part of the work in this subsection.

Towards the end of Subsection 1.1, some important basic properties of probabilities were described. These were that, for any event E, $0 \leq P(E) \leq 1$, with $P(E) = 0$ meaning that event E is impossible and $P(E) = 1$ meaning that event E is certain to happen. These properties of probabilities have important consequences for probability mass functions.

- First, $0 < p(x) \leq 1$ for any value of x in the range of X. This is because $p(x) = P(X = x)$ is the probability of the event '$X = x$'. The reason that $p(x) = 0$ is not allowed is that if any particular value of x is impossible, it is not included in the range of possible values for X.
- Second, since one or other of the values of x in the range of X is sure to happen, the sum of the probabilities of all the possible values is equal to 1; that is, $\sum p(x) = 1$ where the summation is taken over all x in the range of X.

The following box summarises these properties.

Properties of probability mass functions

For a discrete random variable X with probability mass function $p(x)$,

$$0 < p(x) \leq 1$$

for all x in the range of X. Also,

$$\sum p(x) = 1,$$

where the summation is over all x in the range of X.

Example 13 *One is, one isn't*

In Activity 8(b), you obtained the probability mass function associated with rolling a six-sided die on which two faces show a '5' and its other faces show 1, 3, 4 or 6. This p.m.f. is shown in Table 11.

Table 11 The p.m.f. for a die with two faces showing five

x	1	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/3	1/6

This is a valid p.m.f. because $p(x) > 0$ for each x in the range $\{1, 3, 4, 5, 6\}$ and

$$\begin{aligned}\sum p(x) &= p(1) + p(3) + p(4) + p(5) + p(6) \\ &= 1/6 + 1/6 + 1/6 + 1/3 + 1/6 = 1.\end{aligned}$$

Someone else proposed an alternative p.m.f. for another unusual die; it is shown in Table 12.

Table 12 Suggested 'p.m.f.' for die with two faces of five

x	1	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/3	1/3

This is not a valid p.m.f. It does satisfy the first requirement, that $p(x) > 0$ for each x in the range $\{1, 3, 4, 5, 6\}$. However, it does not satisfy the second, that the probabilities add to 1:

$$\begin{aligned}\sum p(x) &= p(1) + p(3) + p(4) + p(5) + p(6) \\ &= 1/6 + 1/6 + 1/6 + 1/3 + 1/3 = 7/6 \neq 1.\end{aligned}$$

Activity 9 *Are they probability mass functions?*

Suppose that X is a random variable with range $\{0, 1, 2, 3\}$. Each of Tables 13–16 purports to be a probability mass function for X. In each case, check whether or not the purported p.m.f. is a valid p.m.f., giving a reason if it is not.

(a) **Table 13** 'P.m.f. 1'

x	0	1	2	3
$p(x)$	0.1	0.4	0.6	−0.1

(b) **Table 14** 'P.m.f. 2'

x	0	1	2	3
$p(x)$	0.1	0.3	0.6	0.1

(c) **Table 15** 'P.m.f. 3'

x	0	1	2	3
$p(x)$	0.1	0.2	0.6	0.1

(d) **Table 16** 'P.m.f. 4'

x	0	1	2	3
$p(x)$	0.3	0.9	−0.3	0

2.3 Probability density functions

Defining a probability function for a *continuous* random variable is a little trickier than for a discrete random variable. It turns out that for continuous random variables, we need a function that can be used to determine probabilities not for a particular value of a random variable but for an interval of values of a random variable. For example, suppose a person's weight is of interest. We require a function that allows us to calculate the probability that the person weighs between, say, 79 kg and 81 kg, or between 62 kg and 66 kg, or even between 71.24 kg and 71.25 kg. The key to forming such a function is to equate 'probability' to 'area', that is, area under a particular curve, and this can be motivated by considering histograms.

Figure 5 (overleaf) shows a frequency histogram of the 100 leaf lengths given in Tables 1 and 3; it is Figure 1 with the values of the (sample) frequencies printed on the histogram boxes. So there were 5 leaves of length 0.0 cm or more and less than 0.5 cm, 20 leaves of length 0.5 cm or more and less than 1.0 cm, and so on.

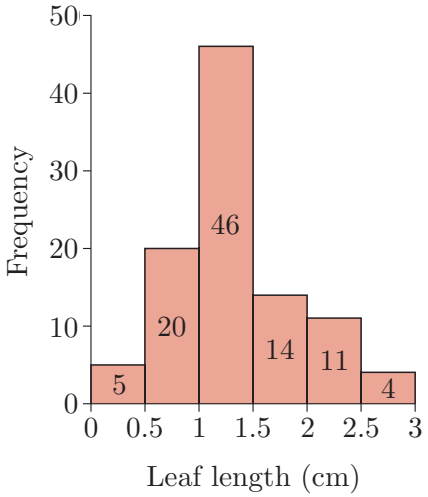


Figure 5 Histogram of leaf lengths with frequencies emphasised

Now, probabilities are equivalent to proportions, which also go by the name of *relative* frequencies. The relative frequencies are the frequencies divided by the total number of items in the sample. The usual, frequency-based, histogram discussed in Unit 1 is produced by making the height of each histogram box equal to the corresponding frequency. In the same way, a *relative frequency histogram* can be produced by making the height of each histogram box equal to the corresponding relative frequency. This is done for the leaf lengths in Figure 6. The numbers printed above the boxes are now the relative frequencies associated with each box. Since there were 100 leaves in the sample, the relative frequency of leaves of length 0.0 cm or more and less than 0.5 cm is $5/100 = 0.05$, the relative frequency of leaves of length 0.5 cm or more and less than 1.0 cm is $20/100 = 0.2$, and so on. Since relative frequencies are proportional to frequencies, the *shape* of the relative frequency histogram in Figure 6 is the same as that of the frequency histogram in Figure 5. It is only the scale and meaning of the vertical axis that has changed.

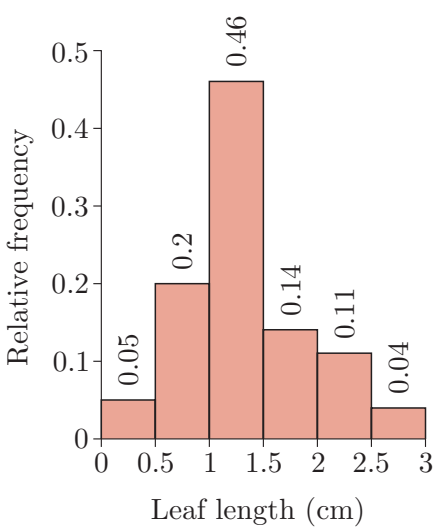


Figure 6 Relative frequency histogram of leaf lengths with relative frequencies emphasised

To this point in this subsection, we have been thinking of frequencies and relative frequencies as being reflected in the heights of histogram boxes. Since the histogram boxes are of equal width, the frequencies and relative frequencies are also *proportional* to the areas of the histogram boxes. For example, the area of the relative frequency histogram box corresponding to leaves of length 0.0 cm or more and less than 0.5 cm is $0.05 \times 0.5 = 0.025$, the area of the relative frequency histogram box corresponding to leaves of length 0.5 cm or more and less than 1.0 cm is $0.2 \times 0.5 = 0.1$, and so on. The total area of all the histogram boxes is

$$0.025 + 0.1 + 0.23 + 0.07 + 0.055 + 0.02 = 0.5.$$

Let us rescale the histogram once again, this time in the same way as we did in Subsection 5.2 of Unit 1, to make the total area of all the histogram boxes equal to one (in Figure 6, the total area is 0.5). To do this, we divide the height of each box by the total area, that is, in this case, divide all the heights by 0.5 (or equivalently, multiply all the heights by 2). The result is a **unit-area histogram**, which was introduced in Unit 1; it is shown for the leaf lengths in Figure 7. Again, the shape of the histogram is unchanged, and indeed so are the relative frequencies, or proportions, printed above the histogram boxes. It is only the vertical scale that has changed: the heights of the boxes are rescaled versions of the relative frequencies. (In the example we looked at in Unit 1, the bin widths were 1, so the relative frequency histogram and the unit-area histogram were the same.)

In more general situations than histograms with equal width boxes, areas and heights are not equivalent, and it turns out to be appropriate to work with areas rather than heights.

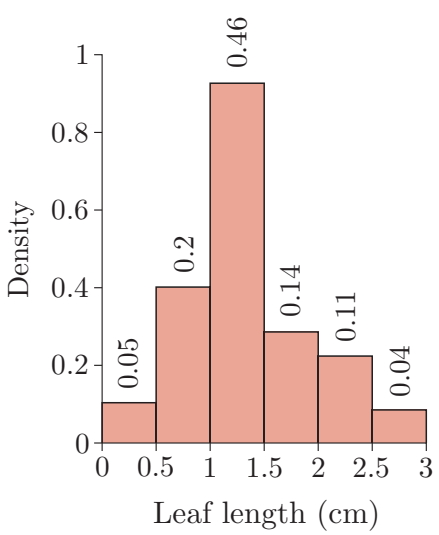


Figure 7 Unit-area histogram of leaf lengths with proportions emphasised

Why go to all this trouble when each histogram has the same shape? Well, on the unit-area histogram we can say that the proportion of the data associated with each box is *equal* to its area (and not just proportional to it). This will be important shortly when we move from histograms to the appropriate probability functions for a continuous random variable.

Suppose we pick one of the hundred leaves at random; let X denote the length of that leaf (in cm). We can read particular proportions and hence probabilities connected with X directly off the unit-area histogram.

Reading from left to right, call the boxes in Figure 7 by the names Box 1, Box 2, ..., Box 6. Then, for example,

$$P(0.5 \leq X < 1.0) = \text{area of Box 2} = 0.2,$$
$$P(0.5 \leq X < 1.5) = \text{area of Box 2} + \text{area of Box 3} = 0.2 + 0.46 = 0.66$$

and

$$P(1.0 \leq X < 2.5) = \text{area of Box 3} + \text{area of Box 4} + \text{area of Box 5}$$
$$= 0.46 + 0.14 + 0.11 = 0.71.$$

Activity 10 *Lengths of scallops*

A dredge survey in Mercury Bay, Whitianga, New Zealand, caught 222 scallops. The lengths of the scallops were measured (in cm) and Figure 8 shows a histogram of the data. The proportions of scallops in each histogram box are written above the box. For instance, the first box shows the proportion of the scallops whose lengths were greater than or equal to 60 cm but less than 70 cm to be 0.225. The proportions are given correct to three decimal places.

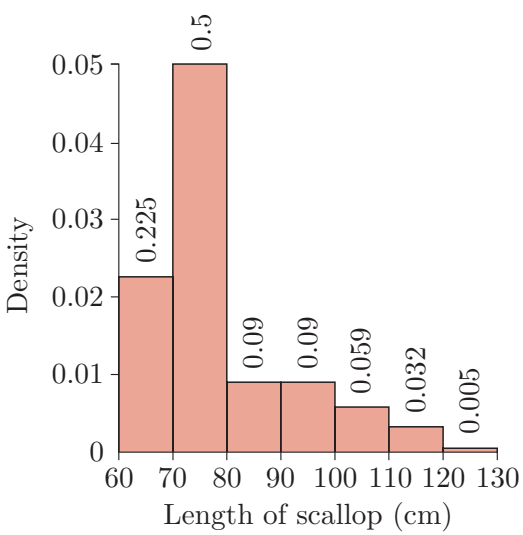


Figure 8 Unit-area histogram of lengths of scallops

(Source: Jorgensen, M.A. (1990) 'Inference-based diagnostics for finite mixture models', *Biometrics*, vol. 46, pp. 1047–58)

(a) Check that the histogram in Figure 8 is a unit-area histogram.

(b) Suppose that one of the 222 scallops is picked at random. Let X denote the length of this randomly chosen scallop (in cm).

(i) What is $P(60 \leq X < 70)$?

(ii) What is $P(70 \leq X < 100)$?

(iii) What is $P(X \geq 90)$?

As discussed earlier, usually the purpose in taking a sample is to learn about the population from which the sample was drawn. With leaf lengths, for example, we would be interested in a larger population of leaves (all the leaves on a particular bush, say, or perhaps all leaves on all such bushes in some region), rather than just the 100 leaves in the sample. Thus probabilities we calculated from a sample would be used as estimates of the corresponding probabilities in the population. For example, 0.66 is both the probability that a randomly chosen leaf from our sample of leaves is between 0.5 cm and 1.5 cm long, and an estimate of the probability that a leaf from the population that gave the sample is between 0.5 cm and 1.5 cm long. But if we took a different sample of size 100, we would not expect the histogram to have precisely the same shape as that in Figure 5, so our estimates of probabilities in the population would change.

However, if we take samples that are large and all the same size, then we should expect their histograms to all be very similar. This is illustrated in Figure 9, which shows histograms for three different samples of leaf lengths. All the samples were of size 1000. (This is just a generalisation of the fact, observed in your computer work in Subsection 2.2, that, for example, the proportion of die rolls producing different outcomes does not vary greatly from sample to sample if large samples are used.)

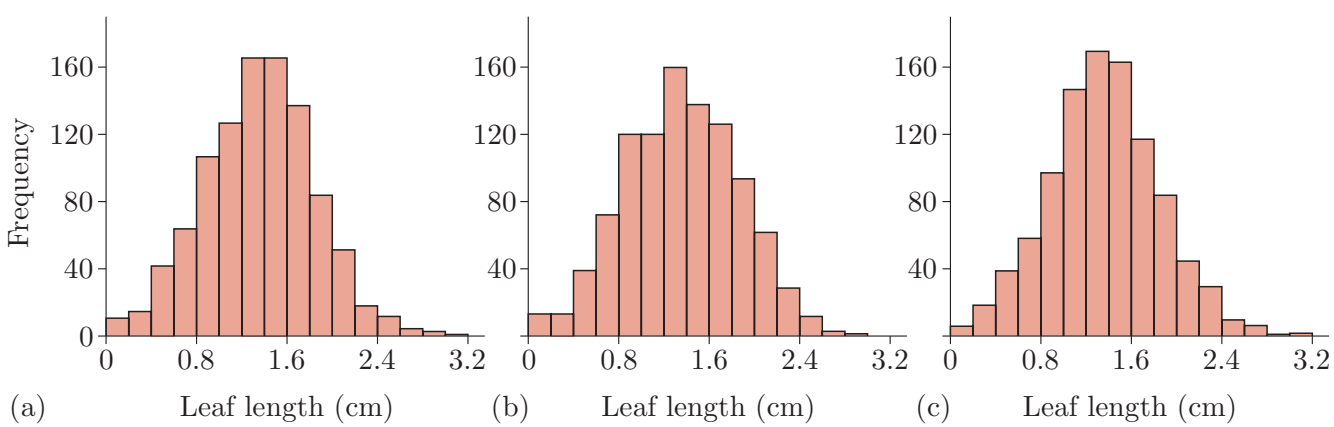


Figure 9 Histograms for three different samples of 1000 leaves

The histograms are very similar to one another. Sample relative frequency estimates of the probability that a randomly selected leaf will be between 0.8 cm and 2.0 cm were obtained for each sample. They were $805/1000 = 0.805$ (sample (a)), $798/1000 = 0.798$ (sample (b)) and $801/1000 = 0.801$ (sample (c)). The fact that these numbers are so close to one another suggests that relative frequencies calculated from large samples provide good estimates of probabilities. In addition, the larger the sample that is taken, the better these estimates are likely to be. Figure 10 (overleaf) shows a histogram based on a very large sample.

The three samples of leaf lengths were in fact generated by computer from a probability model.

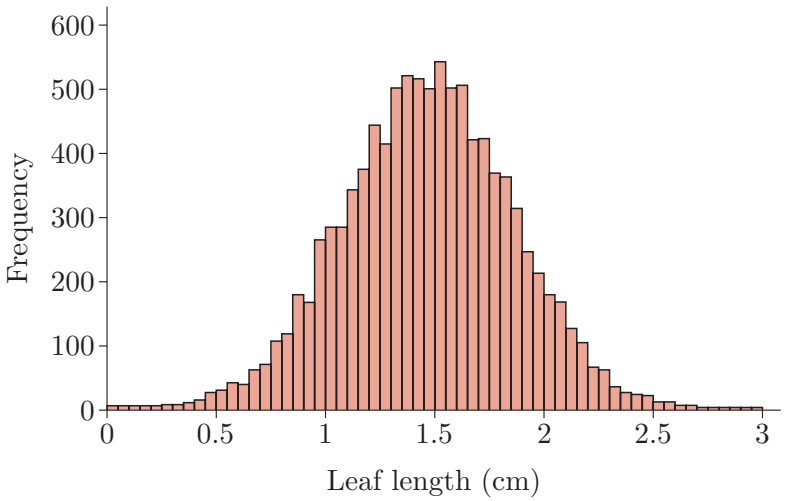


Figure 10 A histogram based on a very large sample

As larger and larger samples are taken, the shapes of the histograms become less jagged, suggesting that a smooth curve might provide an adequate model for the probability distribution of the random variable (see Figure 11).

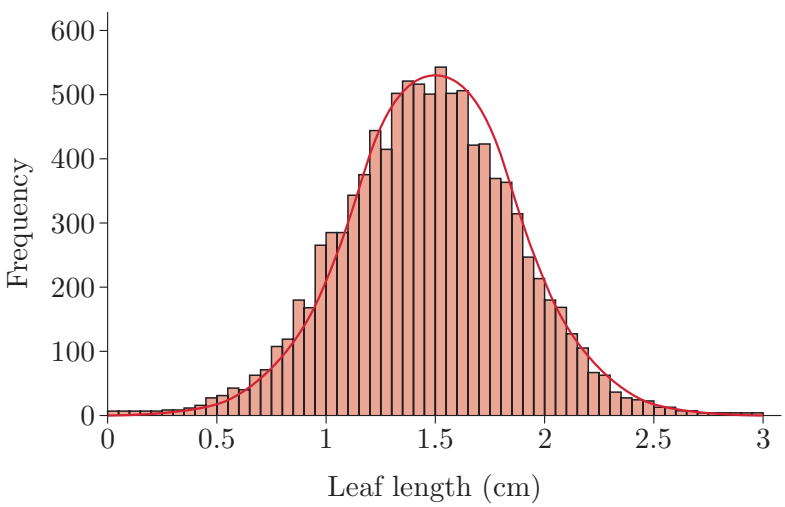


Figure 11 A smooth curve fitted to a histogram

If the curve is scaled, like a unit-area histogram would be, so that the total area under the curve is 1, then, if we wish to know the probability that a randomly plucked leaf will be between 1.0 cm and 1.5 cm (say), we need simply to find the area beneath this curve between 1.0 and 1.5. This is equivalent to finding the total area of the appropriate boxes in a unit-area histogram. This area is shown for the smooth curve of Figure 11 in Figure 12.

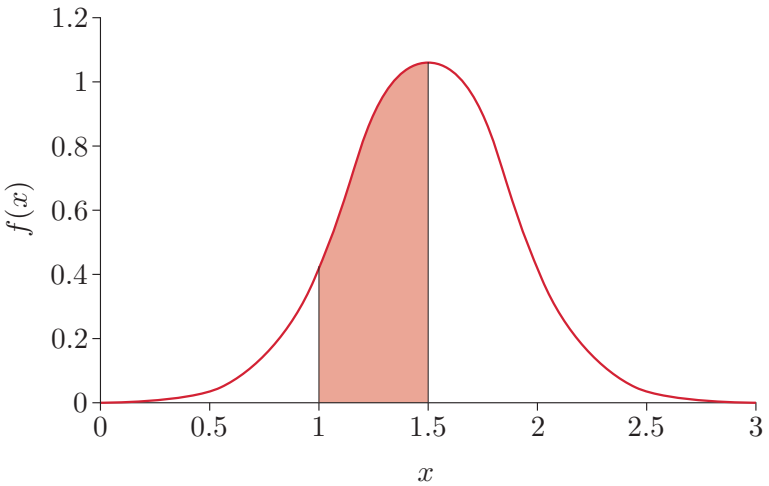


Figure 12 A theoretical probability distribution for leaf lengths

The area of the shaded region in Figure 12 is equal to the probability required. The function which defines the equation of such a curve is called a **probability density function**. The lower-case letter f is commonly used to denote a probability density function, hence the label on the vertical axis in Figure 12. Just as probability mass functions (such as those represented in Figures 3 and 4) provide models for discrete random variables, probability density functions are used to provide models for continuous random variables.

Another example of a sample from a population settling down towards a probability distribution as the sample size increases, only this time for a continuous population, is the topic of Chapter 6 of Computer Book A.

Refer to Chapter 6 of Computer Book A for the next part of the work in this subsection.

In the next activity, you will look at another dataset of continuous measurements and consider the type of probability density function that might be used to model them.

Activity 11 *Traffic data*

The data shown in Table 17 (overleaf) are the 50 time intervals (in seconds) between the first 51 vehicles passing a particular point in one of the lanes of the Kwinana Freeway in Perth, Western Australia, after 9:44 a.m. on a particular day. A histogram for these data is given in Figure 13 (overleaf). The data are recorded as integer numbers of seconds, but a good theoretical model would be continuous: an interval between successive vehicles is extremely unlikely, in reality, to have lasted an exact whole number of seconds.

Kwinana Freeway, Perth, WA

Table 17 Intervals between vehicles, Kwinana Freeway (seconds)

5	8	2	1	8	2	3	5	1	3	1	7	3	3	4	3	4
4	5	2	10	1	5	1	6	14	3	8	5	6	2	5	1	12
6	2	3	2	1	6	7	2	2	4	2	1	1	2	16	2	

(Source: data provided by Professor Toby Lewis)

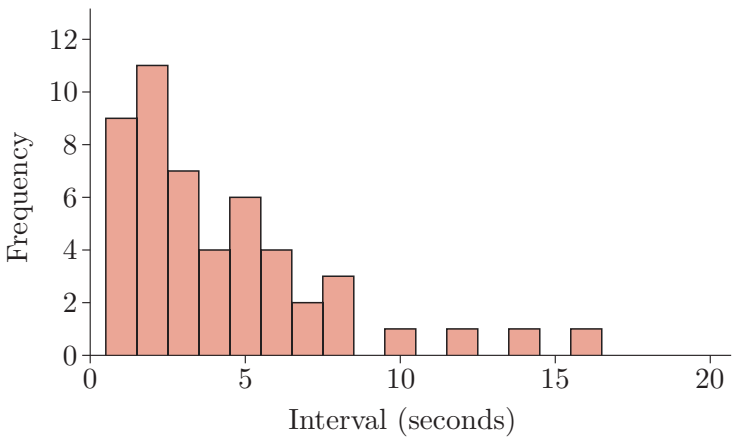


Figure 13 A histogram of the Kwinana Freeway traffic data

You can see from Figure 13 that the random variation exhibited by these time intervals is quite different from that of the leaf lengths. While, for the leaves, intermediate lengths were the most frequent, with shorter and longer measurements occurring less frequently, it appears that for these intervals shorter gaps occur more often than longer ones – the data are very skew.

Sketch a curve which you think might provide a reasonable model for the variation in the lengths of intervals between successive vehicles. Mark on your sketch the area which represents the probability that the length of the interval between two successive vehicles will be between 5 and 10 seconds.

Let us now concentrate on the probability density function, the function that arises as a kind of limiting histogram for huge datasets, and which forms the basis of models for continuous random variables.

The probability density function

For a *continuous* random variable X, observed variation may be modelled by a **probability density function**. This is often abbreviated to **p.d.f.** A probability density function defines a curve, $f(x)$, where f is the standard notation for a p.d.f. The p.d.f. is defined for all values x in the range of X.

The probability that X takes a value between a lower limit x_1 and an upper limit x_2 is then the area under the probability density function $f(x)$ between x_1 and x_2. See Figure 14.

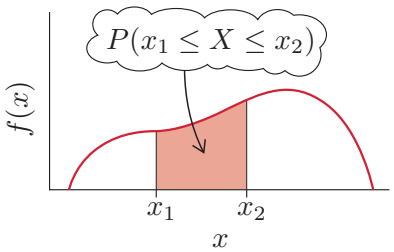


Figure 14 The shaded area is equal to $P(x_1 \leq X \leq x_2)$

Now, recall from your mathematical knowledge that the area under a curve is given by an integral. Therefore the probability that X lies between x_1 and x_2 is the integral of the p.d.f. $f(x)$ between x_1 and x_2, so that

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)\,dx.$$

In this module you are required to perform only simple integrations to calculate such probabilities – you should have learned some calculus before starting M248, but you will be reminded of the results from calculus that we need (and given some revision exercises) before using them in this module.

The important basic properties of probabilities, that $0 \leq P(E) \leq 1$ with $P(E) = 0$ meaning that event E is impossible and $P(E) = 1$ meaning that event E is certain to happen, have consequences for probability density functions as they did for probability mass functions. First, $f(x) \geq 0$ for any value of x in the range of X. Although f is not itself a probability, if $f(x) < 0$ for even a tiny set of values of x, then the probability of X lying in that set would also be negative ... which isn't allowed. Second, since some value of x in the range of X is sure to happen, the total area under the graph of the p.d.f. is equal to 1. The following box summarises these properties.

Properties of probability density functions

For a continuous random variable X with probability density function $f(x)$, the p.d.f. cannot be negative; that is,

$$f(x) \geq 0$$

for all x in the range of X. Also,

$$\int f(x)\,dx = 1,$$

where the integration is over the whole range of possible values of the random variable X. (That is, the total area under the curve defined by the p.d.f. over the entire range of X is equal to 1.)

We will discuss checking that a function purporting to be a probability density function actually has these properties in Subsection 3.3.

Exercises on Section 2

Exercise 3 *Continuous or discrete?*

Customers visit a particular bank on a Monday morning. For each of the following random variables, decide whether a discrete probability model or a continuous probability model would be appropriate.

(a) The number of customers who visit the bank between 10 a.m. and 11 a.m.

(b) The length of time a randomly chosen customer spends in the bank.

(c) The height of a randomly chosen customer.

(d) The number of customers in the queue when a randomly chosen customer enters the bank.

Exercise 4 *The score on unusual dice*

(a) A tetrahedral die has four faces, labelled 1, 2, 3 and 4. The random variable X represents the score on the face on which the die comes to rest when it is rolled. Assuming that the die is unbiased, write down the probability function of X.

(b) An octahedral die has eight faces, labelled $1, 2, \ldots, 8$. The random variable Y represents the score on the face on which the die comes to rest when it is rolled. Assuming that the die is unbiased, write down the probability function of Y.

Exercise 5 *Eruptions of Old Faithful geyser*

In Activity 10(b) of Unit 1, you considered time intervals between eruptions; here you consider the lengths of the eruptions themselves.

Old Faithful still going faithfully

The frequency histogram in Figure 15 represents the durations of 106 eruptions of the Old Faithful geyser in Yellowstone National Park, USA, in August 1978.

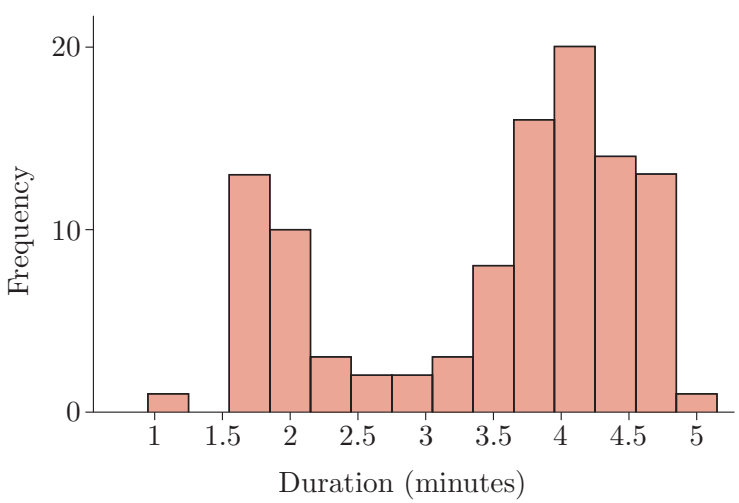


Figure 15 Durations of eruptions of the Old Faithful geyser

(Source: Azzalini, A. and Bowman, A.W. (1990) 'A look at some data on the Old Faithful geyser', *Applied Statistics*, vol. 39, no. 3, pp. 357–65)

(a) Briefly describe the shape of the histogram.

(b) Sketch a curve which you think might reasonably model the variation in the durations of eruptions of the Old Faithful geyser. Mark on your sketch the area which represents the probability that an eruption will last between 3 and 4 minutes.

3 Calculating probabilities in the continuous case

In Section 2, we defined the probability that a continuous random variable X takes a value between limits x_1 and x_2 as the area under the probability density function $f(x)$ between x_1 and x_2. We observed that this area can be calculated using integration.

Calculating probabilities for continuous random variables

For a continuous random variable X with probability density function f, the probability that an observation on X lies between limits x_1 and x_2 may be calculated as

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)\,dx. \tag{1}$$

In Subsection 3.1, we revise some necessary results from calculus. Then, in Subsection 3.2, we use the calculus results in order to obtain probabilities from a p.d.f., using Equation (1). Finally, in Subsection 3.3, we use the calculus results some more to check that functions claimed to be probability density functions really are probability density functions.

This should be revision, rather than meeting integration for the first time.

3.1 Integrating powers and polynomials

In this unit, we will need to integrate quantities like

$$5x^3 \quad \text{and} \quad 5 + 3x - 2x^2 + 4.2x^6.$$

The first quantity is a constant multiple of a *power* of x. The second quantity is a *polynomial function* of x made up of sums of constant multiples of powers of x. To integrate the polynomial you will need to integrate the individual terms in the polynomial and then combine the results. Let's start, then, by integrating the individual terms, which are constants times powers of x.

$\int ax^{-1}\,dx = a\log x$, but this will not be used in this unit.

Suppose we want to integrate ax^k. Assume that $k \neq -1$. The formula is

$$\int ax^k\,dx = \frac{ax^{k+1}}{k+1} + c.$$

In words, we increase the power of x by 1, giving ax^{k+1}, divide by the new power of x, giving $ax^{k+1}/(k+1)$, and finally add a constant c. Notice that the multiplicative constant a remains unchanged.

Two important special cases of this are the integrals of a constant and of x itself.

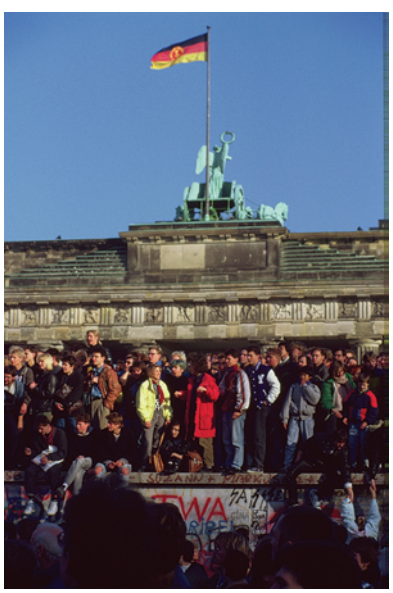
The value of any number raised to the power 0 is 1.

Since $a = ax^0$ and $ax = ax^1$, we have

$$\int a\,dx = ax + c \quad \text{and} \quad \int ax\,dx = \frac{ax^2}{2} + c.$$

Throughout this section the variable of integration will be called 'x' but it doesn't matter what it is called. For example, it is also the case that

$$\int a\,dt = at + c \quad \text{and} \quad \int bz\,dz = \frac{bz^2}{2} + c.$$



The reunification of Germany in 1990; integrating powers?

Example 14 *Integrating powers*

To illustrate applying these rules:

$$\int 7\,dx = 7x + c,$$

$$\int 2x\,dx = \frac{2x^2}{2} + c = x^2 + c,$$

$$\int 5x^3\,dx = \frac{5x^4}{4} + c,$$

$$\int \frac{3}{x^2}\,dx = \int 3x^{-2}\,dx = \frac{3x^{-1}}{-1} + c = -\frac{3}{x} + c$$

and

$$\int 3.1x^{5.2}\,dx = \frac{3.1x^{6.2}}{6.2} + c = \frac{x^{6.2}}{2} + c.$$

Activity 12 *Integrating powers*

Find each of the following integrals.

(a) $\int 6x^2\,dx$ (b) $\int -4\,dx$ (c) $\int 2x^7\,dx$ (d) $\int 3x^{9.1}\,dx$

(e) $\int \frac{2}{x^5}\,dx$ (f) $\int 12x\,dx$

Now that you have been reminded how to integrate constants times powers of x, you also need to know how to combine them to integrate a polynomial function. The rule is straightforward.

If $g(x)$, $h(x)$ and $q(x)$ are any functions of x, then

$$\int \{g(x) + h(x) + \cdots + q(x)\}\, dx = \int g(x)\, dx + \int h(x)\, dx + \cdots + \int q(x)\, dx.$$

That is, to integrate the sum of several terms, simply integrate each term separately and add the integrals together. (The constants of integration are combined into a single constant.)

Example 15 *Integrating a polynomial*

As a first example,

$$\begin{aligned}\int (6 + 3x^2 + 4x^5)\, dx &= \int 6\, dx + \int 3x^2\, dx + \int 4x^5\, dx \\ &= 6x + c_1 + \frac{3x^3}{3} + c_2 + \frac{4x^6}{6} + c_3 \\ &= 6x + x^3 + \frac{2x^6}{3} + c.\end{aligned}$$

Notice that since c_1, c_2 and c_3 are arbitrary constants, so is their sum $c = c_1 + c_2 + c_3$, so we might as well use the latter rather than the former.

Activity 13 *Integrating two polynomials*

Find the following integrals.

(a) $\int (5 + 3x - 2x^2 + 4.2x^6)\, dx$

(b) $\int x(1 + x)^2\, dx$

The first polynomial you are asked to integrate is the one mentioned at the start of this subsection.

Another simple extension of what you have already been doing that is useful more generally is that the integral of a constant times a function is equal to the constant times the integral of the function.

If $g(x)$ is any function of x and a is a constant, then

$$\int a\, g(x)\, dx = a \int g(x)\, dx.$$

Example 16 *Integrating a constant times a polynomial*

In Example 15, we showed that

$$\int (6 + 3x^2 + 4x^5)\,dx = 6x + x^3 + \frac{2x^6}{3} + c.$$

What is the integral of $4(6 + 3x^2 + 4x^5)$? We can avoid integrating this polynomial from scratch by multiplying the result we have for the integral of $6 + 3x^2 + 4x^5$ by 4:

$$\int 4(6 + 3x^2 + 4x^5)\,dx = 4\int (6 + 3x^2 + 4x^5)\,dx = 4\left(6x + x^3 + \frac{2x^6}{3} + c\right).$$

As in Example 15, since c is an arbitrary constant, so is $4c$, which we might as well call c again:

$$\int 4(6 + 3x^2 + 4x^5)\,dx = 4\left(6x + x^3 + \frac{2x^6}{3}\right) + c.$$

And you could multiply out the answer if you wished:

$$\int 4(6 + 3x^2 + 4x^5)\,dx = 24x + 4x^3 + \frac{8x^6}{3} + c.$$

Activity 14 *Integrating a constant times a polynomial*

Use the result of Activity 13(a) to evaluate

$$\int (10 + 6x - 4x^2 + 8.4x^6)\,dx.$$

The integrals considered so far are all *indefinite integrals* because we have not specified a range of values for x over which we are integrating. However, the integrals in which we are most interested have the form

$$\int_{x_1}^{x_2} f(x)\,dx,$$

where $x_1 < x_2$, because this is the formula for $P(x_1 \leq X \leq x_2)$ when $f(x)$ is the probability density function of X. Integrals with specified limits like this are called *definite integrals*.

Definite integral

The quantity

$$\int_{x_1}^{x_2} f(x)\,dx$$

is a *definite integral*. Its value is calculated as follows.

1. Determine the indefinite integral of $f(x)$ but omit the constant c. Call this indefinite integral $G(x)$.
2. Replace x by x_1 in $G(x)$ to give $G(x_1)$. Do the same with x_2 to give $G(x_2)$.
3. Set
$$\int_{x_1}^{x_2} f(x)\,dx = G(x_2) - G(x_1).$$

The constant can be omitted because it cancels out: see Example 17.

The steps in calculating a definite integral are often written as

$$\int_{x_1}^{x_2} f(x)\,dx = \big[G(x)\big]_{x_1}^{x_2} = G(x_2) - G(x_1).$$

(The role of the notation G is explanatory here; you need not explicitly write '$G(x) = \cdots$' in your calculations.)

Example 17 *A definite integral of a power*

This example concerns evaluating the quantity $\int_1^2 9x^2\,dx$. The indefinite integral is

$$\int 9x^2\,dx = \frac{9x^3}{3} + c = 3x^3 + c.$$

Following the rule in the box above, $G(x) = 3x^3$. The required definite integral between $x_1 = 1$ and $x_2 = 2$ is therefore

$$\int_1^2 9x^2\,dx = \big[3x^3\big]_1^2 = (3 \times 2^3) - (3 \times 1^3) = 24 - 3 = 21.$$

What would have happened if we had not omitted the constant c? Well, we'd have

$$\begin{aligned}\int_1^2 9x^2\,dx = \big[3x^3 + c\big]_1^2 &= (3 \times 2^3 + c) - (3 \times 1^3 + c)\\ &= (24 + c) - (3 + c) = 21.\end{aligned}$$

Observe that c cancelled out: it always does this, so it might as well have been omitted in the first place.

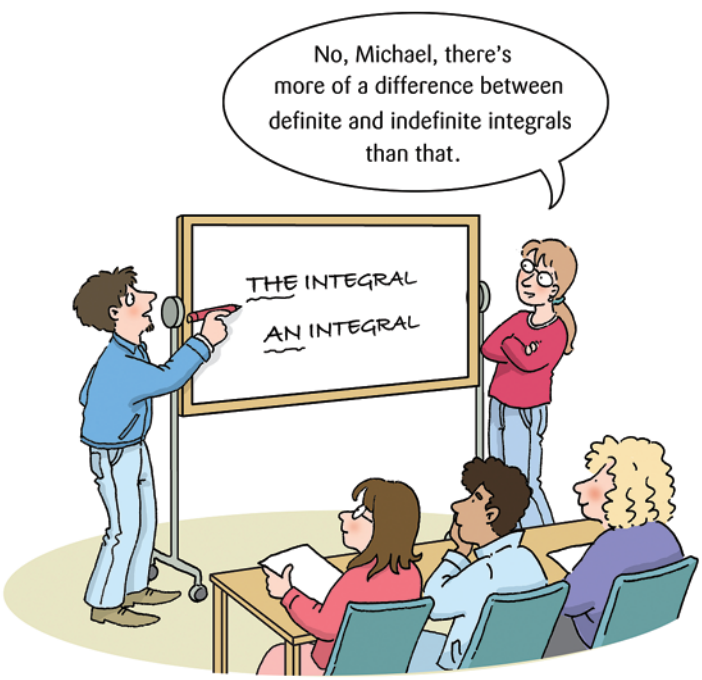
Example 18 *A definite integral of a polynomial*

In this example, we will use the result of Example 15 to calculate a definite integral. We know from Example 15 that

$$\int (6 + 3x^2 + 4x^5)\,dx = 6x + x^3 + \frac{2x^6}{3} + c.$$

We would now like to calculate $\int_1^3 (6 + 3x^2 + 4x^5)\,dx$. This is

$$\begin{aligned}\int_1^3 (6 + 3x^2 + 4x^5)\,dx &= \left[6x + x^3 + \frac{2x^6}{3}\right]_1^3 \\ &= \left(6 \times 3 + 3^3 + \frac{2 \times 3^6}{3}\right) - \left(6 \times 1 + 1^3 + \frac{2 \times 1^6}{3}\right) \\ &= (18 + 27 + 486) - \left(6 + 1 + \frac{2}{3}\right) \\ &= 524 - \frac{2}{3} = \frac{1570}{3} \simeq 523.3.\end{aligned}$$



Definite and indefinite articles

Activity 15 *Definite integrals of a power and two polynomials*

(a) Calculate $\int_{-1}^{1} 3x\,dx$.

(b) Calculate $\int_0^1 x^2(1-2x)\,dx$.

(c) Use the result of Activity 13 to help you calculate

$$\int_0^1 (5+3x-2x^2+4.2x^6)\,dx.$$

If you are unsure about the basic integration methods that you have just worked through, Screencast 2.2 might be of assistance.

Screencast 2.2 Integrating a polynomial

The next subsection will apply your expertise in integration to calculating probabilities associated with probability density functions. The one after that will apply your expertise in integration to checking that a supposed probability density function integrates to 1. Both subsections will give you more practice with integration very similar to that in this subsection.

3.2 Calculating probabilities

Towards the end of Subsection 2.3 you discovered that probabilities could be calculated for continuous probability distributions by integrating probability density functions. In fact, if X is a random variable and f is its p.d.f., then Equation (1) tells us that, for $x_1 < x_2$,

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)\,dx.$$

Example 19 *A power p.d.f.*

Suppose that the random variable X has range $(0,1)$ and that its p.d.f. is given by

$$f(x) = 3x^2, \quad 0 < x < 1.$$

What is the value of $P(1/2 \leq X \leq 3/4)$? This p.d.f. and the required probability are shown in Figure 16.

The probability is the definite integral

$$P\left(\tfrac{1}{2} \leq X \leq \tfrac{3}{4}\right) = \int_{\frac{1}{2}}^{\frac{3}{4}} f(x)\,dx = \int_{\frac{1}{2}}^{\frac{3}{4}} 3x^2\,dx = \left[\frac{3x^3}{3}\right]_{\frac{1}{2}}^{\frac{3}{4}} = \left[x^3\right]_{\frac{1}{2}}^{\frac{3}{4}}$$
$$= \left(\frac{3}{4}\right)^3 - \left(\frac{1}{2}\right)^3 = \frac{27}{64} - \frac{1}{8} = \frac{19}{64} \simeq 0.297.$$

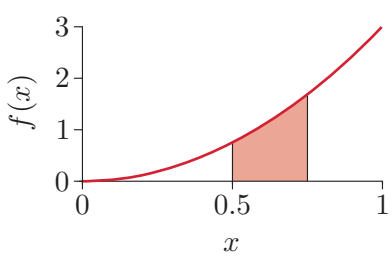


Figure 16 The p.d.f. (curve) and probability (shaded area)

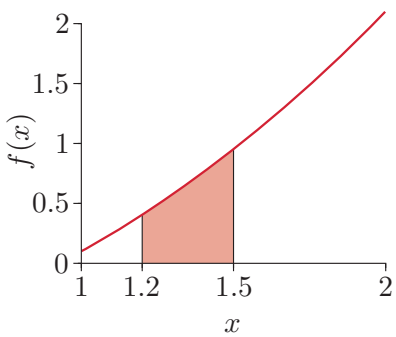


Figure 17 The p.d.f. (curve) and probability (shaded area)

Example 20 *A polynomial p.d.f.*

Suppose that the random variable X has range $(1, 2)$ and that its p.d.f. is given by

$$f(x) = 0.6x^2 + 0.2x - 0.7, \quad 1 < x < 2.$$

We want to calculate $P(1.2 \leq X \leq 1.5)$. This p.d.f. and the required probability are shown in Figure 17.

We have

$$\begin{aligned} P(1.2 \leq X \leq 1.5) &= \int_{1.2}^{1.5} \left(0.6x^2 + 0.2x - 0.7\right) dx \\ &= \left[0.6\,\frac{x^3}{3} + 0.2\,\frac{x^2}{2} - 0.7x\right]_{1.2}^{1.5} \\ &= \left[0.2x^3 + 0.1x^2 - 0.7x\right]_{1.2}^{1.5} \\ &= 0.2(1.5)^3 + 0.1(1.5)^2 - 0.7(1.5) \\ &\quad - \{\,0.2(1.2)^3 + 0.1(1.2)^2 - 0.7(1.2)\} \\ &= 0.2004. \end{aligned}$$

Activity 16 *A probability from a power p.d.f.*

Suppose that the range of the continuous random variable X is from 5 to 10, and within that range its p.d.f. is $f(x) = 10/x^2$. This p.d.f. is shown in Figure 18.

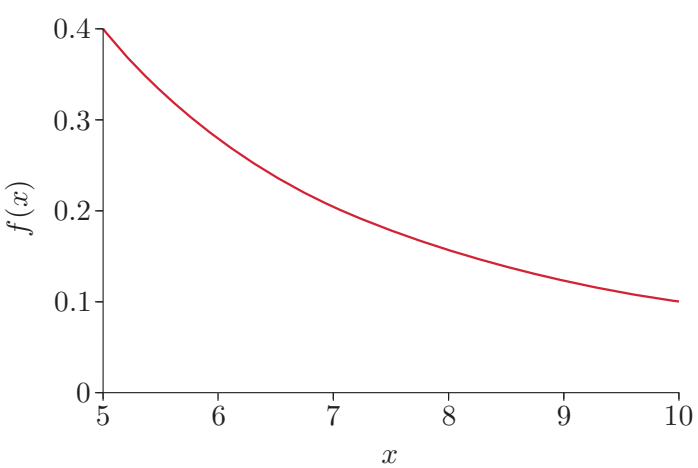


Figure 18 The p.d.f.

(a) On a sketch of Figure 18, shade in the area associated with the probability $P(7 \leq X \leq 8)$.

(b) Calculate $P(7 \leq X \leq 8)$.

Activity 17 *Journey time*

A man's journey to work takes between 20 and 30 minutes. His journey time turns out to be well represented by a random variable X whose p.d.f. is

$$f(x) = \frac{1}{5} - \frac{x}{250}, \quad 20 < x < 30.$$

This p.d.f., which is a linear function of x, is shown in Figure 19.

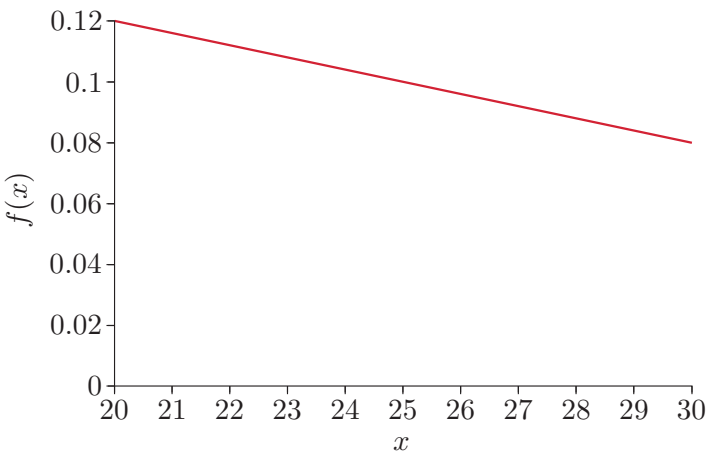


Figure 19 The p.d.f.

(a) On a sketch of Figure 19, shade in the area associated with the probability $P(X > 25)$.

(b) Calculate $P(X > 25)$.

The link between probabilities and integrals leads us to a rather remarkable fact, one that especially distinguishes the continuous case from the discrete case.

Given any particular value, the probability that a continuous random variable takes precisely that value is considered to be zero.

You need not worry in this module about the deeper mathematics behind this result. You should accept the fact because of how the probability of a continuous random variable lying in an interval decreases to zero as the interval gets shorter and shorter. Let $\varepsilon > 0$. It should at least seem plausible to you that as $\varepsilon \to 0$,

$$P(x - \varepsilon \leq X \leq x + \varepsilon) \to P(X = x)$$

and that

$$P(x - \varepsilon \leq X \leq x + \varepsilon) = \int_{x-\varepsilon}^{x+\varepsilon} f(x)\,dx \to 0.$$

ε is the Greek lower-case letter epsilon

The symbol '$\to$' is read as 'tends to'.

You do *not* need to be able to prove these results.

Happily, this apparently arcane property will actually serve to simplify some further probability calculations in Subsection 4.3.

3.3 Checking that a function is a probability density function

In Subsection 3.2, it has several times been stated that a particular function is a probability density function. We should really check some of these claims! The properties required of a function f in order for it to be a valid p.d.f. were set out at the end of Section 2. They are that:

- $f(x) \geq 0$ for all x in the range of X
- $\int f(x)\,dx = 1$, where the integration is over the whole range of possible values of the random variable X.

The non-negativity requirement is often quite easily checked by simple mathematics or by sketching the function; in this module, we will ask you to consider the non-negativity – or otherwise – of only very simple functions.

This is very general: L could be $-\infty$ and/or U could be ∞ as well as one or both of them being finite (with $L < U$).

Let the range of X be written as (L, U). Then the integration requirement can be written

$$\int_L^U f(x)\,dx = 1.$$

Example 21 *Validity of a polynomial p.d.f.*

In Example 20, we presumed that

$$f(x) = 0.6x^2 + 0.2x - 0.7, \quad 1 < x < 2,$$

is a probability density function. But is this really so? Well, f can be shown to be positive over the range of x mathematically (which you needn't bother with here) or pictorially, as in Figure 17. But does f integrate to 1 over its range? Since the range of f is $(1, 2)$, we need to evaluate

$$\begin{aligned}\int_1^2 (0.6x^2 + 0.2x - 0.7)\,dx &= \left[0.2x^3 + 0.1x^2 - 0.7x\right]_1^2 \\ &= 0.2(2)^3 + 0.1(2)^2 - 0.7(2) \\ &\quad - \left\{0.2(1)^3 + 0.1(1)^2 - 0.7(1)\right\} \\ &= 1.6 + 0.4 - 1.4 - (0.2 + 0.1 - 0.7) = 1.\end{aligned}$$

So yes, f is a valid probability distribution function.

Example 22 *Validity of the journey time p.d.f.*

In Activity 17, we presumed that

$$f(x) = \frac{1}{5} - \frac{x}{250}, \quad 20 < x < 30,$$

is a valid p.d.f. This f is a linear function, so it will be non-negative for all

x in the range if it is non-negative at each end of the range. This is so:

$$f(20) = \frac{1}{5} - \frac{20}{250} = \frac{3}{25} \quad \text{and} \quad f(30) = \frac{1}{5} - \frac{30}{250} = \frac{2}{25},$$

which are both positive. (Alternatively, f can be seen to be positive over its range from Figure 19.)

The integral of f over its range is

$$\int_{20}^{30} \left(\frac{1}{5} - \frac{x}{250}\right) dx = \left[\frac{x}{5} - \frac{x^2}{500}\right]_{20}^{30} = \frac{30}{5} - \frac{900}{500} - \left(\frac{20}{5} - \frac{400}{500}\right)$$
$$= 6 - 1.8 - 4 + 0.8 = 1.$$

So f does integrate to 1 over its range, and f is a valid probability density function.

Activity 18 *Validity of two power p.d.f.s*

(a) In Example 19, we presumed that

$$f(x) = 3x^2, \quad 0 < x < 1,$$

is a valid p.d.f. Check the two properties of probability density functions that confirm that this is so.

(b) In Activity 16, we presumed that

$$f(x) = \frac{10}{x^2}, \quad 5 < x < 10,$$

is a valid p.d.f. Check that this is the case.

A non-negative function, g say, is sometimes suggested as the p.d.f. of a probability distribution but the function doesn't integrate to 1 over its range. In such cases, we can adapt the function so that it *will* integrate to 1 over its range, and the adapted function will be a valid p.d.f. Suppose, therefore, that $g(x) \geq 0$ for $L < x < U$, but that

$$\int_L^U g(x)\,dx = K, \quad K > 0.$$

Then if we define

$$f(x) = \frac{1}{K}\,g(x),$$

it is the case that

$$\int_L^U f(x)\,dx = \int_L^U \frac{1}{K}\,g(x)\,dx = \frac{1}{K}\int_L^U g(x)\,dx = \frac{1}{K}\,K = 1.$$

This always works unless $\int_L^U g(x)\,dx = \infty$.

Thus $f(x)$ is non-negative and *does* integrate to 1 over (L, U), and so is a valid p.d.f. The constant K is often called the *normalising constant* because it makes the function integrate to 1.

Some statisticians call $1/K$ the normalising constant.

Example 23 *Making a valid p.d.f*

A botanist is interested in the values of the proportions of the surface area of the flowers of a particular orchid that are coloured red. He considers the proportions to be values of a random variable X with range $0 < X < 1$ and suggests that a good model for his data would have a p.d.f. that varied with x as

$$g(x) = 1 - x, \quad 0 < x < 1.$$

This is a non-negative function because it is linear over its range, taking non-negative values $g(0) = 1$ and $g(1) = 0$ at the ends of the range.

But is g a valid p.d.f.: does it integrate to 1 over its range? Well,

$$\int_0^1 (1 - x)\, dx = \left[x - \frac{x^2}{2}\right]_0^1 = 1 - \frac{1}{2} - (0 - 0) = \frac{1}{2} \neq 1.$$

So g is not a valid p.d.f. However, if we set

$$K = \int_0^1 (1 - x)\, dx = \frac{1}{2},$$

then

$$f(x) = \frac{1}{K}\, g(x) = \frac{1}{1/2}(1 - x) = 2(1 - x), \quad 0 < x < 1,$$

is a valid probability density function.

Activity 19 *Making more valid p.d.f.s*

(a) The function

$$g(x) = 9x^2, \quad 1 < x < 2,$$

is non-negative (for all x and hence in particular for $1 < x < 2$). It is not, however, a valid p.d.f. because we showed in Example 17 that $\int_1^2 9x^2\, dx = 21$. By introducing a suitable normalising constant, obtain a valid p.d.f. that is proportional to g.

(b) The function

$$g(x) = (x - 1)^2, \quad 1 < x < 6,$$

is non-negative (for all x, because it is a squared quantity, and hence in particular for $1 < x < 6$). By introducing a suitable normalising constant, obtain a valid p.d.f. that is proportional to g.

Screencast 2.3 reiterates what makes a function a probability density function and works through another example. Screencast 2.3 uses the indefinite integral that was obtained in Screencast 2.2.

Screencast 2.3 Functions that are probability density functions

Exercises on Section 3

Exercise 6 *Practice with integration*

Evaluate the following integrals.

(a) $\displaystyle\int_{-1}^{2} x^3(1-x)\,dx$ (b) $\displaystyle\int_{1}^{2}\left(3x^2-\frac{1}{x^2}\right)dx$

Exercise 7 *Practice with probabilities*

(a) Suppose that the random variable X has range $(0,1)$ and that its p.d.f. is given by

$$f(x) = 2(1-x), \quad 0 < x < 1.$$

What is the value of $P\left(\frac{1}{4} \leq X \leq \frac{1}{2}\right)$?

(b) Suppose that the random variable X has range $(1,6)$ and that its p.d.f. is given by

$$f(x) = \frac{3}{125}(x-1)^2, \quad 1 < x < 6.$$

What is the value of $P(2 \leq X \leq 5)$?

Exercise 8 *Are they probability density functions?*

Suppose that X is a random variable with range $(0,1)$. Each of the following functions purports to be a probability density function for X. In each case, check whether or not the function is a valid p.d.f., giving a reason if it is not.

(a) 'P.d.f. 1': $f(x) = 2x - 1$

(b) 'P.d.f. 2': $f(x) = \dfrac{1}{Kx^2}$ for an appropriate value of the constant $K > 0$

(c) 'P.d.f. 3': $f(x) = \frac{3}{2}(2-x)$

(d) 'P.d.f. 4': $f(x) = \frac{2}{3}(2-x)$

4 Cumulative distribution functions

Using the appropriate probability function, we can calculate the probability that a random variable will lie in any given interval by summing probabilities if the random variable is discrete, or by integration if the random variable is continuous. However, there is another function associated with a probability distribution which often simplifies such calculations, especially when a number of probabilities are to be computed. This function – the *cumulative distribution function* – is closely related to probability mass functions and probability density functions. It is defined simultaneously for both discrete and continuous random variables in

Subsection 4.1. Subsections 4.2 and 4.3 then consider the cumulative distribution function in more detail in the discrete case and in the continuous case, respectively.

4.1 The cumulative distribution function in general

Suppose that we wish to find the probability that a random variable X will not exceed some specified value x. That is, we want to find the probability $P(X \leq x)$.

If X is discrete, then this probability may be obtained from the p.m.f. $p(x)$ by summing appropriate terms: if X has range $\{0, 1, 2, \ldots\}$, for instance, then this probability may be written as

$$P(X \leq x) = \sum_{j=0}^{x} p(j) = p(0) + p(1) + \cdots + p(x).$$

On the other hand, if X is continuous, then this probability may be obtained from the p.d.f. $f(x)$ by finding an appropriate area under the graph of the p.d.f.: for example, if X takes only non-negative values, then $P(X \leq x) = P(0 \leq X \leq x)$, so this probability may be written as

$$P(X \leq x) = \int_0^x f(y)\, dy.$$

See Subsection 4.3 for discussion of why $f(y)\, dy$ has been use in the integral, rather than $f(x)\, dx$.

Whether a random variable X is discrete or continuous, the function F defined by $F(x) = P(X \leq x)$ is called the **cumulative distribution function** of the random variable X. The notation $F(.)$ is standard for a cumulative distribution function, for both discrete and continuous random variables. Note that the cumulative distribution function F is defined for *any* random variable X, discrete or continuous, whatever its range.

The cumulative distribution function

The **cumulative distribution function** F of a random variable X is a function which, for each value x in the range of X, gives the probability that X takes a value less than or equal to x:

$$F(x) = P(X \leq x).$$

The simpler term **distribution function** is sometimes used for the cumulative distribution function. The abbreviation **c.d.f.** is frequently used.

4.2 The cumulative distribution function in the discrete case

In the following example, the c.d.f. of a particular discrete random variable is obtained.

Example 24 *The c.d.f. of the score on an unbiased die*

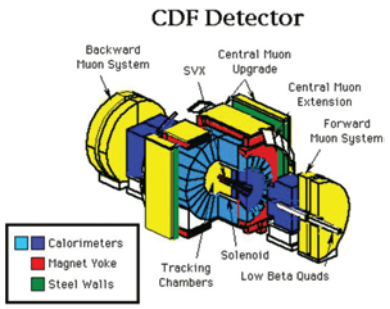
For an unbiased die, the probability mass function of Y, the score that appears when it is rolled, is given by

$$p(y) = 1/6, \quad y = 1, 2, 3, 4, 5, 6.$$

(See Example 11.) The cumulative distribution function is defined by $F(y) = P(Y \leq y)$. So, for instance,

$$F(3) = P(Y \leq 3) = P(Y = 1 \text{ or } 2 \text{ or } 3)$$
$$= p(1) + p(2) + p(3) = 3/6 = 1/2.$$

Values of the p.m.f. $p(y)$, and of the c.d.f. $F(y)$ of the random variable Y obtained by addition from the p.m.f., are given in Table 18. A table such as this is a convenient way of setting out values of the p.m.f. and the c.d.f. The p.m.f. was visualised in Figure 3; this is repeated in Figure 20(a) alongside the c.d.f. in Figure 20(b).



You won't be needing one of these machines from particle physics for this example!

Table 18 The probability distribution for the score on an unbiased die

y	1	2	3	4	5	6
$p(y)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$F(y)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{5}{6}$	1

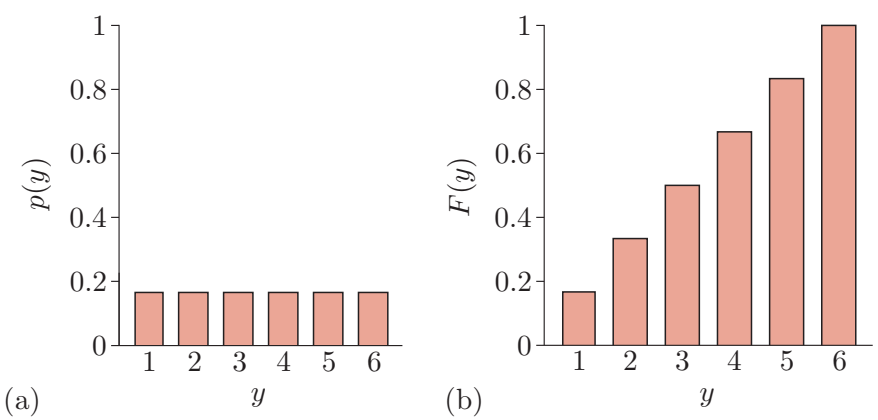


Figure 20 (a) The p.m.f. and (b) the c.d.f. for an unbiased die

The c.d.f. can be used directly to write down particular probabilities. For example, using the values listed in the table, the probability that the score obtained when an unbiased die is rolled is at most 2 is

$$P(Y \leq 2) = F(2) = \tfrac{1}{3}.$$

The probability that the score is less than 5 is

$$P(Y < 5) = P(Y \leq 4) = F(4) = \tfrac{2}{3};$$

here, the equivalence between the events $\{Y < 5\}$ and $\{Y \leq 4\}$ has been used.

The probability that the score is greater than 4 is

$$P(Y > 4) = 1 - P(Y \leq 4) = 1 - F(4) = 1 - \tfrac{2}{3} = \tfrac{1}{3}.$$

This is an application of the probability rule for complementary events mentioned in Section 1.

To obtain this result, notice that the events $\{Y > 4\}$ and $\{Y \leq 4\}$ are complementary, that is, one or other of them must happen. Therefore $P(Y > 4) + P(Y \leq 4) = 1$ and hence $P(Y > 4) = 1 - P(Y \leq 4)$.

Of course, all the probabilities in this example could quite easily have been calculated directly from the p.m.f. However, in more complicated situations, particularly when using the p.m.f. would involve summing a large number of probabilities, it is often easier to use the c.d.f. if it is available. The next two activities will give you some practice at obtaining the c.d.f. of a random variable and using it to find probabilities.

Activity 20 *The c.d.f. for the score on a die with two faces showing five*

See Activity 8(b).

The random variable X, which takes the values $x = 1, 3, 4, 5, 6$, is used to model the outcome of a roll of a die that has a '5' on two faces and its other faces show 1, 3, 4 or 6. (That is, the two-spot face has been replaced by a second five-spot face.)

(a) Construct a table similar to Table 18 to display values of the p.m.f. and c.d.f. of X.

(b) Use the values of the c.d.f. in your table to write down the probability that the score obtained when the die is rolled is as follows.

(i) At most 4 (ii) Less than 4 (iii) Greater than 3

Activity 21 *Occupants of cars*

This one with 21 occupants didn't arrive at the service station!

The number of occupants in a car arriving at a particular service station is a random variable X. Values of the probability mass function of X are given in Table 19.

Table 19 A probability distribution for the number of occupants in a car

x	1	2	3	4	5
$p(x)$	0.4	0.2	0.2	0.1	0.1

(a) Construct a table similar to Table 18 to display values of the p.m.f. and the c.d.f. of X.

(b) Use the values of the c.d.f. in your table to write down the probability that the number of occupants in a car arriving at the service station will be as follows.

(i) At most 1 (ii) Less than 3 (iii) Greater than 2

(iv) At least 4

4.3 The cumulative distribution function in the continuous case

When X is a continuous random variable, its cumulative distribution function is still defined as

$$F(x) = P(X \leq x),$$

but now this probability is an integral.

Cumulative distribution function for a continuous random variable

Suppose that X is a continuous random variable whose range of possible values has a lower limit of L and an upper limit of U. If its p.d.f. is $f(x)$, then its cumulative distribution function is

$$F(x) = \int_L^x f(y)\,dy \quad \text{for } L < x < U.$$

When L and U are finite, a more complete description of the c.d.f. is given by

$$F(x) = \begin{cases} 0 & x \leq L \\ \displaystyle\int_L^x f(y)\,dy & L < x < U \\ 1 & x \geq U. \end{cases}$$

This reflects the facts that values of X less than L are impossible and that values of X *less than values of* x *which are greater than* U are certain to occur (because values greater than U can't). In this module, we will not be fussy, and so will take the version of $F(x)$ in the box as a proxy for the version of $F(x)$ that takes up three lines.

As with the probabilities in Subsection 3.2, $F(x) = P(X \leq x)$ is a shaded area under the p.d.f. f, as shown in Figure 21(a). As this probability is a function of a single x, it can be graphed as a function of x; this is done in Figure 21(b).

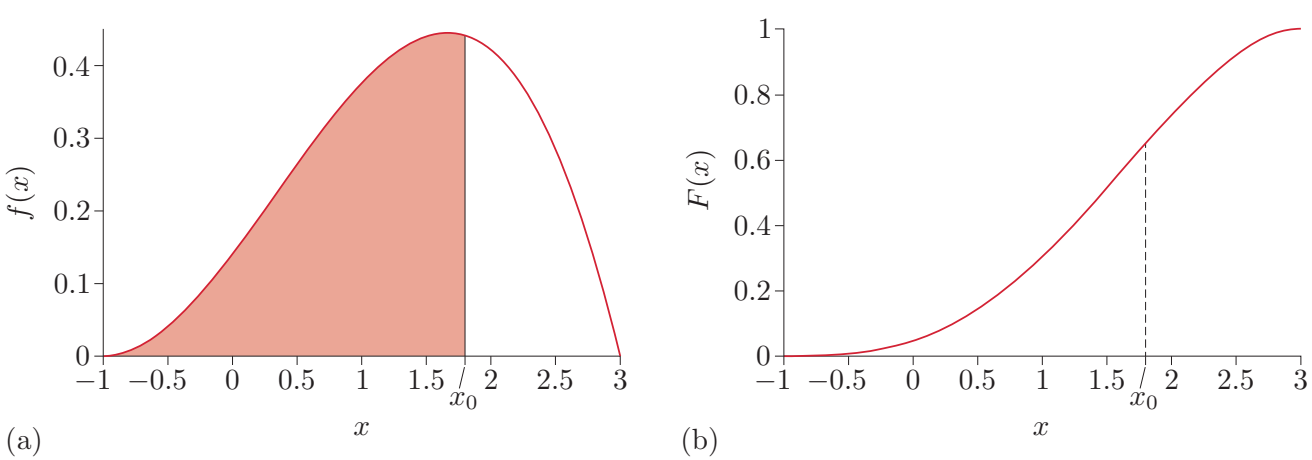


Figure 21 (a) Shaded area under a p.d.f. showing $F(x_0) = P(X \leq x_0)$, (b) $F(x)$ as a function of x

The c.d.f. in Figure 21(b) shows the main properties of any c.d.f. when X is continuous. As well as starting from a value of 0 and ending at a value of 1, we have the following.

In the continuous case, the c.d.f. $F(x)$ is an *increasing* function of x.

The easiest way to see this is by further consideration of graphs like those in Figure 21; see the following screencast for explanation.

Screencast 2.4 Cumulative distribution functions are increasing

Returning to the mathematics, another important point is that we have written $F(x) = \int_L^x f(y)\,dy$ not $F(x) = \int_L^x f(x)\,dx$. In Section 3, we said that what we used as the variable of integration doesn't matter ... but that was when the limits of integration were numbers. However, now we are interested in the c.d.f. at the value x, and hence need to use x as a limit of integration. It is, therefore, important not to confuse yourself (or others) by calling the variable of integration and a limit of integration the same things. (The variable of integration y could, of course, have been z or t or anything else that is not x.)

Example 25 *A power c.d.f.*

In Example 19, we considered the random variable X with range $(0, 1)$ and p.d.f. given by

$$f(x) = 3x^2, \quad 0 < x < 1.$$

What is the c.d.f. associated with X?

Here, $L = 0$. So for $0 < x < 1$,

$$F(x) = \int_0^x f(y)\,dy = \int_0^x 3y^2\,dy = \left[y^3\right]_0^x = x^3 - 0 = x^3.$$

The c.d.f. is shown in Figure 22.

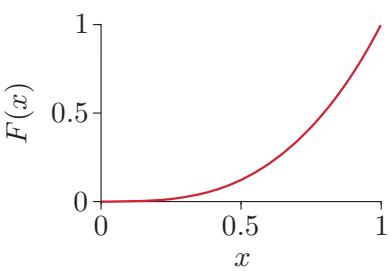


Figure 22 The power c.d.f.

Example 26 *A polynomial c.d.f.*

In Example 20, we considered the random variable X with range $(1, 2)$ and p.d.f. given by

$$f(x) = 0.6x^2 + 0.2x - 0.7, \quad 1 < x < 2.$$

Here, $L = 1$. The c.d.f. is, for $1 < x < 2$, given by

$$\begin{aligned} F(x) &= \int_1^x (0.6y^2 + 0.2y - 0.7)\,dy = \left[0.2y^3 + 0.1y^2 - 0.7y\right]_1^x \\ &= 0.2x^3 + 0.1x^2 - 0.7x - (0.2 + 0.1 - 0.7) \\ &= 0.2x^3 + 0.1x^2 - 0.7x + 0.4. \end{aligned}$$

The c.d.f. is shown in Figure 23.

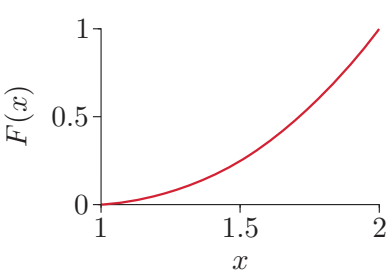


Figure 23 The polynomial c.d.f.

Activity 22 *A power c.d.f.*

In Activity 16, you considered the random variable X with range $(5, 10)$ and p.d.f. given by

$$f(x) = 10/x^2, \quad 5 < x < 10.$$

Find the c.d.f.

Activity 23 *C.d.f. of journey time*

In Activity 17, you considered the random variable X representing a man's journey time to work. X has range $(20, 30)$ minutes and p.d.f.

$$f(x) = \frac{1}{5} - \frac{x}{250}, \quad 20 < x < 30.$$

Find the c.d.f.

You have seen that the c.d.f. is defined in the same way for discrete and for continuous random variables. However, it is important to distinguish between the two types of random variables when using the c.d.f. to calculate probabilities. For example, if X is continuous, then the probabilities $P(X < x)$ and $P(X \leq x)$ may both be represented by the area under the p.d.f. of X to the left of x; that is,

$$P(X \leq x) = P(X < x) = F(x).$$

This is because the probability that X takes exactly the value x is effectively zero in the continuous case (see the end of Subsection 3.2). On the other hand, for a discrete random variable Y, the probabilities $P(Y < y)$ and $P(Y \leq y)$ are *not* in general equal. For instance, if Y is the score obtained when an unbiased six-sided die is rolled, then

See Table 18.

$$P(Y < 4) = P(Y \leq 3) = F(3) = 1/2$$

but

$$P(Y \leq 4) = F(4) = 2/3.$$

So, in this case, $P(Y < 4) \neq P(Y \leq 4)$. The difference between the two is $P(Y = 4) = p(4) = 1/6 \neq 0$.

Concentrating again on the continuous case, applications of the probability rule for complementary events mentioned in Section 1 give that

$$P(X \geq x) + P(X < x) = 1 \quad \text{and} \quad P(X > x) + P(X \leq x) = 1,$$

so

$$P(X \geq x) = 1 - P(X < x) \quad \text{and} \quad P(X > x) = 1 - P(X \leq x).$$

But

$$P(X \leq x) = P(X < x) = F(x),$$

so

$$P(X \geq x) = P(X > x) = 1 - F(x).$$

In Subsection 3.2, we were concerned with probabilities of the form $P(x_1 \leq X \leq x_2)$ where $x_1 < x_2$. As above, we can now recognise that *in the continuous case*, the same probability value ensues for probabilities of each of the forms $P(x_1 \leq X < x_2)$, $P(x_1 < X \leq x_2)$ and $P(x_1 < X < x_2)$ also! Whichever version of these probabilities you want, the formula is the same.

Reverting to the initial version of this probability for concreteness, we would like to know how it, $P(x_1 \leq X \leq x_2)$, relates to the c.d.f. of X. The answer is readily seen from Figure 24. The shaded area under the p.d.f. in Figure 24(a) is $F(x_2)$; the shaded area under the p.d.f. in Figure 24(b) is $F(x_1)$; and the shaded area under the p.d.f. in Figure 24(c) is $P(x_1 \leq X \leq x_2)$. It is clear, however, that the shaded area in Figure 24(c) is equal to the shaded area in Figure 24(a) *minus* the shaded area in Figure 24(b). In symbols, we have shown that

In integrals, $\int_{x_1}^{x_2} f(x)\,dx = \int_L^{x_2} f(x)\,dx - \int_L^{x_1} f(x)\,dx$.

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1).$$

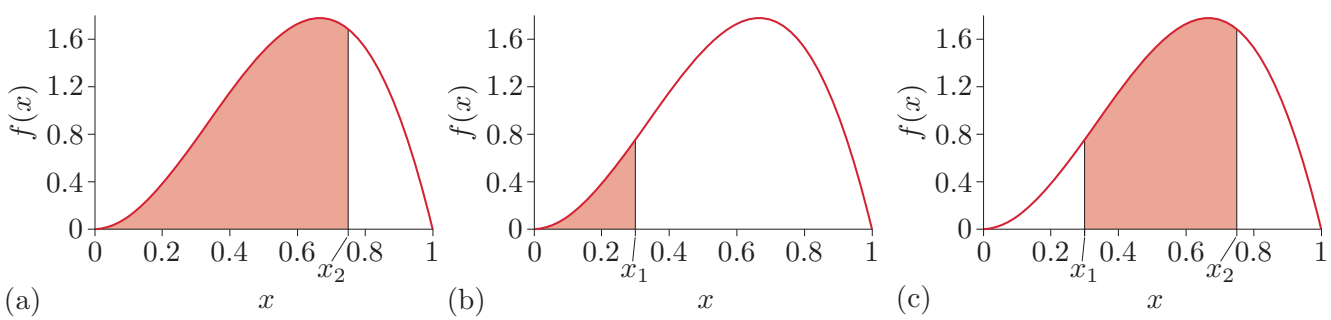


Figure 24 Shaded areas under a p.d.f. showing (a) $F(x_2)$, (b) $F(x_1)$, (c) $P(x_1 \leq X \leq x_2)$

It is worth highlighting this important relationship between probabilities and cumulative distribution functions.

Probabilities of lying within intervals and c.d.f.s

For continuous X,

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1). \qquad (2)$$

Note again that the highlighted relationship and those in the paragraphs above it are for continuous distributions only. *They do not, in general, hold for discrete distributions.*

We can now use Equation (2) to simplify calculations of probabilities of lying within intervals for some continuous distributions. It especially

comes into its own by saving you from having to do repeated integrations when several probabilities are required.

Example 27 *A probability from a polynomial c.d.f.*

In Examples 20 and 26, we considered the random variable X with range $(1, 2)$ and p.d.f. given by

$$f(x) = 0.6x^2 + 0.2x - 0.7, \quad 1 < x < 2.$$

In Example 26, we showed that its c.d.f. is

$$F(x) = 0.2x^3 + 0.1x^2 - 0.7x + 0.4, \quad 1 < x < 2.$$

Using Equation (2), we find that

$$\begin{aligned} P(1.2 \leq X \leq 1.5) &= F(1.5) - F(1.2) \\ &= 0.2(1.5)^3 + 0.1(1.5)^2 - 0.7(1.5) + 0.4 \\ &\quad - \left\{0.2(1.2)^3 + 0.1(1.2)^2 - 0.7(1.2) + 0.4\right\} \\ &= 0.2004, \end{aligned}$$

hence confirming the result from Example 20.

You might well object that if we take into account the integration by which we calculated $F(x)$ in the first place, nothing has really been saved or simplified. This is true. But now consider working out another probability for the same distribution, say $P(X \geq 1.75)$. No further integration is required because we have the formula for $F(x)$. So we have

$$\begin{aligned} P(X \geq 1.75) &= 1 - F(1.75) \\ &= 1 - \left\{0.2(1.75)^3 + 0.1(1.75)^2 - 0.7(1.75) + 0.4\right\} \\ &\simeq 0.447. \end{aligned}$$

Activity 24 *Probabilities from c.d.f.s*

(a) In Examples 19 and 25, we considered the random variable X with range $(0, 1)$ and p.d.f. given by

$$f(x) = 3x^2, \quad 0 < x < 1.$$

In Example 25, we showed that its c.d.f. is

$$F(x) = x^3, \quad 0 < x < 1.$$

(i) Using Equation (2), confirm the result from Example 19 that $P(1/2 \leq X \leq 3/4) = 19/64$.

(ii) What is $P(X > 0.6)$?

(iii) What is $P(0.1 < X \leq 0.6)$?

(b) In Activities 17 and 23, you considered the random variable X representing a man's journey time to work. X has range $(20, 30)$ minutes and p.d.f.

$$f(x) = \frac{1}{5} - \frac{x}{250}, \quad 20 < x < 30.$$

You showed in Activity 23 that its c.d.f. is

$$F(x) = \frac{x}{5} - \frac{x^2}{500} - \frac{16}{5}, \quad 20 < x < 30.$$

(i) What is the probability that the man's journey time is greater than 22 minutes?

(ii) What is the probability that the man's journey time is between 21 and 29 minutes?

Screencast 2.5 works through calculation of the c.d.f. and of probabilities therefrom for the p.d.f. developed in Screencast 2.3.

Screencast 2.5 Cumulative distribution function and probabilities

The final activity in this section, and hence this unit, gives you further practice in each of the aspects of probability models for continuous data that you have learned about in this unit. Some of its manipulations are a little harder than in the examples and activities so far: don't spend a long time on it if you get bogged down in the detail.

Activity 25 *Bulldozer return times*

Model based on data in AbouRizk, S.M., Halpin, D.W. and Wilson, J.R. (1994) 'Fitting beta distributions based on sample data', *Journal of Construction Engineering and Management*, vol. 120, no. 2, pp. 288–305.

A study was made of the times taken for a bulldozer to complete a particular task as part of earthmoving operations. These 'return times' (in minutes) were centred at a little less than one minute but could take up to two. An estimation exercise of the kind you will be considering later in the module suggested a model for the distribution of bulldozer return times, X, having p.d.f. given by

$$f(x) = \frac{15}{16\sqrt{2}}\sqrt{x}(2-x), \quad 0 < x < 2.$$

This p.d.f. is shown in Figure 25.

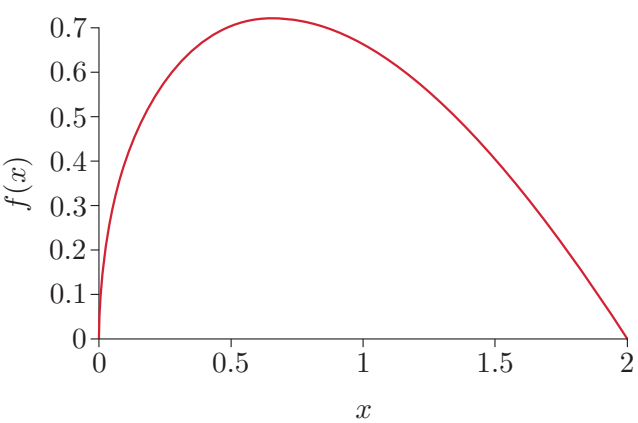


Figure 25 The bulldozer p.d.f.

(a) Verify that f is a valid p.d.f.

(b) Show that the c.d.f. is given by

$$F(x) = \frac{1}{8\sqrt{2}}\, x\sqrt{x}(10 - 3x).$$

(c) What is the probability that the bulldozer's return time is greater than a minute?

(d) What is the probability that the bulldozer's return time is between 30 seconds and a minute?

Exercises on Section 4

Exercise 9 *The score on an octahedral die*

An octahedral die has eight faces labelled $1, 2, \ldots, 8$. The random variable Y represents the score on the face on which the die comes to rest when it is rolled. Assume that the die is unbiased.

(a) Construct a table similar to Table 18 to display values of the p.m.f. and the c.d.f. of Y.

(b) Use the c.d.f. to write down the probability that the score on the die will be as follows.

(i) At most 3 (ii) Less than 6 (iii) Greater than 4

(iv) At least 4

Exercise 10 *Length of brown trout fry*

Brown trout are bred in a hatchery pond and sold, primarily for release to the wild, according to size. The smallest brown trout 'fry' that are sold are 3–6 cm in length. Within a batch of such fry, the lengths (in cm) can be approximated by a random variable X whose p.d.f. is

$$f(x) = \frac{1}{30}\left(10x - x^2 - 14\right) \quad \text{for } 3 < x < 6.$$

This p.d.f. is shown in Figure 26 (overleaf).

(a) Verify that f is a valid p.d.f. You may assume that the function $10x - x^2 - 14$ is non-negative for $3 < x < 6$; see Figure 26.

(b) What is the c.d.f., F, associated with f?

(c) What is the probability that a randomly chosen fry from the batch is less than 4 cm long?

(d) What is the probability that a randomly chosen fry from the batch is between 4 cm and 5 cm long?

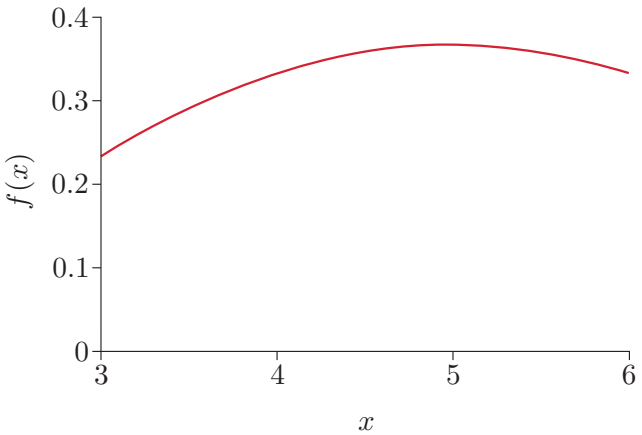


Figure 26 The trout fry p.d.f.

Summary

In this unit, you have been introduced to some basic ideas about modelling the variation observed in a sample of data. A fundamental idea is that of using a number – a probability – to measure how likely a chance event is to occur. You have also met the notion of a random variable, and the need for two essentially different types of models for random variables has been discussed: discrete models for counts and other discrete data, and continuous models for measurements. You have seen how a probability model may be specified using a probability function: a probability mass function for a discrete model and a probability density function for a continuous model. The cumulative distribution function of a probability distribution has also been defined. This function is very useful for calculating probabilities from models, especially when several such probabilities are required.

In the case of models for continuous data, you have used integration to calculate probabilities, to check that a function claimed to be a probability density function really is one, and to calculate cumulative distribution functions.

Learning outcomes

After you have worked through this unit, you should be able to:

- appreciate that a probability is a number between 0 and 1 (inclusive)
- estimate a probability, given data
- calculate a probability when assumptions about the symmetry of an object or situation may be made
- appreciate the 'settling down' phenomenon which occurs when a statistical experiment is repeated many times
- understand how probabilities of outcomes are encapsulated in the probability mass function (p.m.f.) of models for discrete data
- calculate probabilities of outcomes using the probability density function (p.d.f.) of models for continuous data
- appreciate that the total area under the graph of a p.d.f. is equal to 1
- check that functions purporting to be p.m.f.s and p.d.f.s are valid
- understand the definition of the cumulative distribution function (c.d.f.)
- in simple situations, write down the p.m.f. and the c.d.f. of a discrete random variable and use these to calculate probabilities
- use the c.d.f. of a continuous random variable to calculate probabilities of lying within intervals
- use integration to obtain results associated with continuous probability distributions that have simple p.d.f.s.

Solutions to activities

Solution to Activity 1

(a) Assuming that the ball is equally likely to come to rest in any of the 37 compartments, the probability that it will come to rest in any particular compartment is 1/37. So the probability that the ball will come to rest in the compartment numbered 19 is 1/37.

(b) Of the 37 compartments, 18 are odd-numbered $(1, 3, 5, \ldots, 35)$, so the probability that the ball will come to rest in an odd-numbered compartment is 18/37.

Solution to Activity 2

The proportion of colour blind pupils is $2/25 = 0.08$. So if a pupil is picked at random from the class, the probability that the pupil is colour blind is 0.08.

Solution to Activity 3

The probability that an adult living in the UK has outstanding credit card debts is estimated by the proportion of adults in the survey who had outstanding credit card debts. This is $474/2000 \simeq 0.24$.

Solution to Activity 4

(a) An estimate of the probability that a male will be given help is

$$\frac{71}{71+29} = \frac{71}{100} = 0.71.$$

(b) An estimate of the probability that a female will be given help is

$$\frac{89}{89+16} = \frac{89}{105} \simeq 0.85.$$

(c) Since 0.85 is greater than 0.71, the experiment has provided some evidence to support the notion that people are more helpful to females than to males. However, two questions arise. First, is the difference between the observed proportions sufficiently large to indicate a genuine difference in helping behaviour, or could it have arisen simply as a consequence of experimental variation when in fact there is no underlying difference in people's willingness to help others, whether male or female? Second, is the design of the experiment adequate to answer the research question? There may have been differences (other than their sexes) between the eight students that influenced people's responses. One matter not addressed in this activity, but surely relevant to the investigation, is the sex of those approached.

Solution to Activity 5

The sample relative frequency for an event E is the proportion of times that the event occurs, so it is always a number between 0 and 1. Since a probability is the value towards which the sample relative frequency tends as the number of repetitions of an experiment increases, it must also be a number between 0 and 1.

Solution to Activity 6

The range is $\{1, 2, 3, 4, 5, 6\}$.

Solution to Activity 7

(a) The data have been obtained by *measuring* the lengths of kangaroos' jawbones. Evidently the lengths have been recorded to the nearest 0.1 mm, but the actual lengths of kangaroo jawbones are not restricted in this way – within a reasonable range, any length is possible. The random variable is continuous.

(b) A *count* of yeast cells in each square is bound to result in an integer observation: you could not have 2.2 or 3.4 cells. The random variable is discrete.

(c) The coding of a 'remain' vote to $Y = 1$ and a 'leave' vote to $Y = 0$ has made Y a discrete random variable (which happens to be binary).

(d) The failure times have been *measured* to the nearest 0.1 minute and recorded as such. However, failure time is a continuous random variable: components would not fail only at tenths of a minute. A useful model would be a continuous model.

(e) A *count* of spontaneous fission tracks in each grain is bound to result in an integer observation: you could not have 1.88 or 101.125 tracks. The random variable is discrete.

Solution to Activity 8

(a) Let X be 1 if a randomly chosen street light from the consignment is faulty, and let X be 0 if a randomly chosen street light from the consignment is not faulty. Its probability mass function can be written as

$$p(x) = \begin{cases} 0.96 & x = 0 \\ 0.04 & x = 1. \end{cases}$$

Of course, you do not have to call the random variable X, nor do you have to assign these particular two numerical values '1' and '0' to 'faulty' and 'not faulty', respectively (but you do have to assign unequal numerical values to the two cases).

(b) As two of the six faces give a five, $p(5) = 2/6 = 1/3$, while each of the other possible outcomes has a probability of $1/6$ of occurring. The probability mass function might be written as

$$p(x) = \begin{cases} 1/3 & x = 5 \\ 1/6 & x = 1, 3, 4, 6 \end{cases}$$

or as

Table 20

x	1	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/3	1/6

Solution to Activity 9

(a) 'P.m.f. 1' is not a valid p.m.f. because $p(3) = -0.1 < 0$.

(b) 'P.m.f. 2' is not a valid p.m.f. because $\sum p(x) = 1.1 > 1$.

(c) 'P.m.f. 3' is a valid p.m.f.: $0 < p(x) \leq 1, x = 0, 1, 2, 3$, and $\sum p(x) = 1$.

(d) 'P.m.f. 4' is not a valid p.m.f. for three reasons: $p(2) = -0.3 < 0$; $p(3) = 0$; and $\sum p(x) = 0.9$.

Solution to Activity 10

(a) The total area of the histogram boxes is

$$0.225 + 0.5 + 0.09 + 0.09 + 0.059 + 0.032 + 0.005 = 1.001.$$

It seems that this is a unit-area histogram, assuming that the value 1.001, rather than 1, arises as a result of rounding error (the proportions are not exact but given correct to three decimal places).

(b) Reading from left to right, call the boxes in Figure 8 by the names Box 1, Box 2, ..., Box 7.

(i) $P(60 \leq X < 70) =$ area of Box 1 $= 0.225$.

(ii) $P(70 \leq X < 100) =$ area of Box 2 + area of Box 3 + area of Box 4
$= 0.5 + 0.09 + 0.09 = 0.68$.

(iii) $P(X \geq 90) =$ area of Box 4 + area of Box 5
+ area of Box 6 + area of Box 7
$= 0.09 + 0.059 + 0.032 + 0.005 = 0.186$.

Solution to Activity 11

The sample is not a large one and the histogram is quite jagged, so there is not a clear-cut answer to what the shape of the curve should be. However, since the data are very skew, the curve should also be skew. One possibility is shown in Figure 27. Notice that time intervals cannot be negative, so the probability density function should start from zero (that is, the range of this distribution has a lower limit of zero). Assuming that the curve has been scaled so that the total area under the curve is equal to 1, the shaded area represents the probability that the interval between two successive vehicles will be between 5 and 10 seconds.

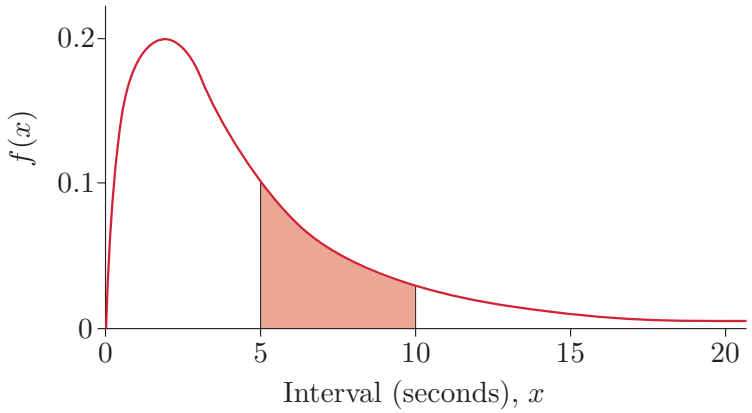


Figure 27 A possible model for intervals between vehicles

Solution to Activity 12

(a) $\displaystyle\int 6x^2\,dx = \frac{6x^3}{3} + c = 2x^3 + c.$

(b) $\displaystyle\int -4\,dx = -4x + c.$

(c) $\displaystyle\int 2x^7\,dx = \frac{2x^8}{8} + c = \frac{1}{4}x^8 + c.$

(d) $\displaystyle\int 3x^{9.1}\,dx = \frac{3x^{10.1}}{10.1} + c \simeq 0.297x^{10.1} + c.$

(e) $\displaystyle\int \frac{2}{x^5}\,dx = \int 2x^{-5}\,dx = \frac{2x^{-4}}{-4} + c = -\frac{1}{2x^4} + c.$

(f) $\displaystyle\int 12x\,dx = \frac{12x^2}{2} + c = 6x^2 + c.$

Solution to Activity 13

(a) $$\int (5 + 3x - 2x^2 + 4.2x^6)\,dx = 5x + \frac{3x^2}{2} - \frac{2x^3}{3} + \frac{4.2x^7}{7} + c$$
$$= 5x + \frac{3x^2}{2} - \frac{2x^3}{3} + \frac{3x^7}{5} + c.$$

(b) Before integrating, we need to multiply out the function:

$$\int x(1+x)^2\,dx = \int x(1 + 2x + x^2)\,dx = \int (x + 2x^2 + x^3)\,dx$$
$$= \frac{x^2}{2} + \frac{2x^3}{3} + \frac{x^4}{4} + c.$$

Solution to Activity 14

Because $10 + 6x - 4x^2 + 8.4x^6 = 2(5 + 3x - 2x^2 + 4.2x^6)$, we have

$$\begin{aligned}\int (10 + 6x - 4x^2 + 8.4x^6)\,dx &= 2\int (5 + 3x - 2x^2 + 4.2x^6)\,dx \\ &= 2\left(5x + \frac{3x^2}{2} - \frac{2x^3}{3} + \frac{3x^7}{5}\right) + c \\ &= 10x + 3x^2 - \frac{4x^3}{3} + \frac{6x^7}{5} + c.\end{aligned}$$

Solution to Activity 15

(a) $$\int_{-1}^{1} 3x\,dx = \left[\frac{3x^2}{2}\right]_{-1}^{1} = \frac{3\times 1^2}{2} - \frac{3\times(-1)^2}{2} = \frac{3}{2} - \frac{3}{2} = 0.$$

(b) $$\begin{aligned}\int_0^1 x^2(1-2x)\,dx &= \int_0^1 (x^2 - 2x^3)\,dx = \left[\frac{x^3}{3} - \frac{2x^4}{4}\right]_0^1 \\ &= \left(\frac{1^3}{3} - \frac{2\times 1^4}{4}\right) - \left(\frac{0^3}{3} - \frac{2\times 0^4}{4}\right) = \left(\frac{1}{3} - \frac{1}{2}\right) - 0 = -\frac{1}{6}.\end{aligned}$$

(c) $$\begin{aligned}\int_0^1 (5 + 3x - 2x^2 + 4.2x^6)\,dx &= \left[5x + \frac{3x^2}{2} - \frac{2x^3}{3} + \frac{3x^7}{5}\right]_0^1 \\ &= \left(5\times 1 + \frac{3\times 1^2}{2} - \frac{2\times 1^3}{3} + \frac{3\times 1^7}{5}\right) \\ &\quad - \left(5\times 0 + \frac{3\times 0^2}{2} - \frac{2\times 0^3}{3} + \frac{3\times 0^7}{5}\right) \\ &= \left(5 + \frac{3}{2} - \frac{2}{3} + \frac{3}{5}\right) - 0 \\ &= 5 + \frac{43}{30} = \frac{193}{30} \simeq 6.433.\end{aligned}$$

Solution to Activity 16

(a) The required probability is shown in Figure 28.

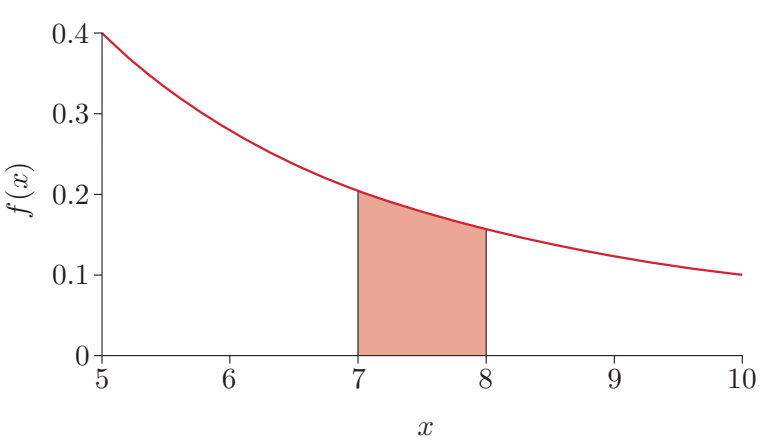


Figure 28 The probability is the shaded area

(b) $P(7 \leq X \leq 8) = \displaystyle\int_7^8 10x^{-2}\,dx = \left[10\,\frac{x^{-1}}{-1}\right]_7^8 = \left[-\frac{10}{x}\right]_7^8$

$$= -\frac{10}{8} - \left(-\frac{10}{7}\right) = \frac{10}{56} = \frac{5}{28} \simeq 0.179.$$

Solution to Activity 17

(a) Notice that for this distribution, $P(X > 25) = P(25 \leq X \leq 30)$. The required probability is therefore as shown in Figure 29.

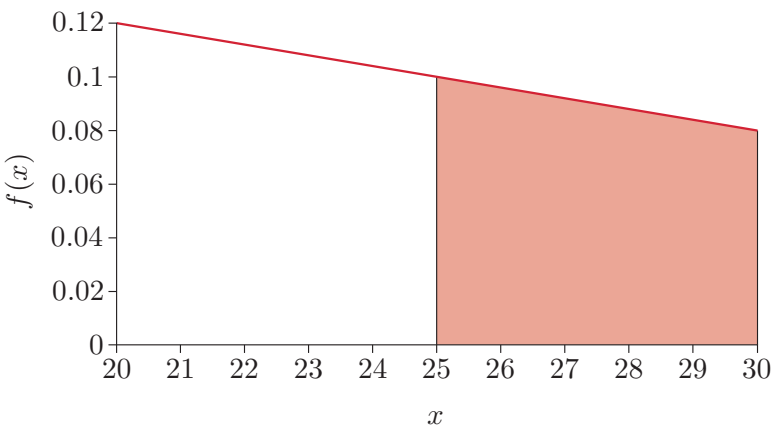


Figure 29 The probability is the shaded area

(b) $P(X > 25) = P(25 \leq X \leq 30)$

$$= \int_{25}^{30} \left(\frac{1}{5} - \frac{x}{250}\right) dx = \left[\frac{x}{5} - \frac{x^2}{500}\right]_{25}^{30}$$

$$= \frac{30}{5} - \frac{900}{500} - \left(\frac{25}{5} - \frac{625}{500}\right)$$

$$= 6 - 1.8 - 5 + 1.25 = 0.45.$$

Solution to Activity 18

(a) The function $3x^2$, being a positive constant times a squared quantity, is non-negative for all x and so it is, in particular, non-negative for $0 < x < 1$.

The integral of f over its range is

$$\int_0^1 3x^2\,dx = \left[x^3\right]_0^1 = 1^3 - 0^3 = 1.$$

So f is a valid p.d.f.

(b) The function $10/x^2$, being a positive constant divided by a squared quantity, is also non-negative for all x and so it is, in particular, non-negative for $5 < x < 10$.

The integral of f over its range is

$$\int_5^{10} 10x^{-2}\,dx = \left[-\frac{10}{x}\right]_5^{10} = -1-(-2) = 1.$$

So f is a valid p.d.f.

Solution to Activity 19

(a) The normalising constant is $K = \int_1^2 9x^2\,dx = 21$, so

$$f(x) = \frac{1}{K}\,9x^2 = \frac{9}{21}x^2 = \frac{3}{7}x^2, \quad 1 < x < 2,$$

is a valid p.d.f. proportional to g.

(b) The normalising constant is

$$K = \int_1^6 (x-1)^2\,dx = \int_1^6 (x^2-2x+1)\,dx = \left[\frac{x^3}{3} - x^2 + x\right]_1^6$$
$$= 72 - 36 + 6 - \left(\frac{1}{3} - 1 + 1\right) = \frac{125}{3}.$$

Thus

$$f(x) = \frac{1}{K}\,(x-1)^2 = \frac{3}{125}(x-1)^2, \quad 1 < x < 6,$$

is a valid p.d.f. proportional to g.

Solution to Activity 20

(a) In Activity 8(b), you found the probability mass function $p(x)$ for the score on a die whose '2' had been replaced with a '5'. This is shown in the following table, together with the cumulative distribution function $F(x)$. The c.d.f. was found by summing values of the p.m.f.; for example,

$$P(X \leq 5) = p(1) + p(3) + p(4) + p(5) = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{3} = \tfrac{5}{6}.$$

Table 21

x	1	3	4	5	6
$p(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$
$F(x)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{5}{6}$	1

(b) (i) $P(X \leq 4) = F(4) = \frac{1}{2}$.

(ii) $P(X < 4) = P(X \leq 3) = F(3) = \frac{1}{3}$.

(iii) $P(X > 3) = 1 - P(X \leq 3) = 1 - F(3) = 1 - \frac{1}{3} = \frac{2}{3}$.

Solution to Activity 21

(a) Values of the c.d.f. are included in the following table. They were obtained by summing values of the p.m.f.

Table 22

x	1	2	3	4	5
$p(x)$	0.4	0.2	0.2	0.1	0.1
$F(x)$	0.4	0.6	0.8	0.9	1

(b) (i) $P(X \leq 1) = F(1) = 0.4$.

(ii) $P(X < 3) = P(X \leq 2) = F(2) = 0.6$.

(iii) $P(X > 2) = 1 - P(X \leq 2) = 1 - F(2) = 1 - 0.6 = 0.4$.

(iv) $P(X \geq 4) = 1 - P(X \leq 3) = 1 - F(3) = 1 - 0.8 = 0.2$.

Solution to Activity 22

For $5 < x < 10$,

$$F(x) = \int_5^x 10y^{-2}\,dy = \left[-\frac{10}{y}\right]_5^x = \left\{-\frac{10}{x} - \left(-\frac{10}{5}\right)\right\} = 2 - \frac{10}{x}.$$

Solution to Activity 23

For $20 < x < 30$,

$$F(x) = \int_{20}^x \left(\frac{1}{5} - \frac{y}{250}\right) dy = \left[\frac{y}{5} - \frac{y^2}{500}\right]_{20}^x$$
$$= \frac{x}{5} - \frac{x^2}{500} - \left(\frac{20}{5} - \frac{400}{500}\right) = \frac{x}{5} - \frac{x^2}{500} - \frac{16}{5}.$$

Solution to Activity 24

(a) (i) $P(1/2 \leq X \leq 3/4) = F(3/4) - F(1/2)$

$$= \left(\frac{3}{4}\right)^3 - \left(\frac{1}{2}\right)^3 = \frac{27}{64} - \frac{1}{8} = \frac{19}{64},$$

as was to be confirmed.

(ii) $P(X > 0.6) = 1 - F(0.6) = 1 - (0.6)^3 = 0.784$.

(iii) $P(0.1 < X \leq 0.6) = F(0.6) - F(0.1) = (0.6)^3 - (0.1)^3 = 0.215$.

(b) (i) $P(X > 22) = 1 - F(22) = 1 - \left(\frac{22}{5} - \frac{22^2}{500} - \frac{16}{5}\right) = 0.768$.

(ii) $P(21 \leq X \leq 29) = F(29) - F(21)$

$$= \frac{29}{5} - \frac{29^2}{500} - \frac{16}{5} - \left(\frac{21}{5} - \frac{21^2}{500} - \frac{16}{5}\right) = 0.8.$$

Solution to Activity 25

(a) f is non-negative because for $0 < x < 2$, the constant, the square root term and the linear term are all non-negative. (The linear term, $2 - x$, decreases from 2 when $x = 0$ to 0 when $x = 2$.)

Also, $\int_0^2 f(x)\,dx = 1$. To see this,

$$\begin{aligned}\int_0^2 \frac{15}{16\sqrt{2}}\sqrt{x}(2-x)\,dx &= \frac{15}{16\sqrt{2}}\int_0^2 (2x^{1/2} - x^{3/2})\,dx \\ &= \frac{15}{16\sqrt{2}}\left[\frac{4x^{3/2}}{3} - \frac{2x^{5/2}}{5}\right]_0^2 \\ &= \frac{15}{16\sqrt{2}}\left(\frac{4\times 2\times\sqrt{2}}{3} - \frac{2\times 4\times\sqrt{2}}{5} - (0-0)\right) \\ &= \frac{15}{16\sqrt{2}} \times 8\sqrt{2} \times \left(\frac{1}{3} - \frac{1}{5}\right) = \frac{15}{2}\times\frac{2}{15} = 1.\end{aligned}$$

Therefore f is a valid p.d.f.

(b) For $0 < x < 2$,

$$\begin{aligned}F(x) &= \int_0^x \frac{15}{16\sqrt{2}}\sqrt{y}(2-y)\,dy = \frac{15}{16\sqrt{2}}\left[\frac{4y^{3/2}}{3} - \frac{2y^{5/2}}{5}\right]_0^x \\ &= \frac{15}{16\sqrt{2}}\left(\frac{4x\sqrt{x}}{3} - \frac{2x^2\sqrt{x}}{5} - (0-0)\right) \\ &= \frac{15}{16\sqrt{2}} \times 2x\sqrt{x} \times \left(\frac{2}{3} - \frac{x}{5}\right) \\ &= \frac{15}{8\sqrt{2}} \times x\sqrt{x} \times \frac{1}{15}(10 - 3x) = \frac{1}{8\sqrt{2}}\,x\sqrt{x}(10-3x),\end{aligned}$$

as required.

(c) You are asked for $P(X > 1)$. This is

$$P(X > 1) = 1 - F(1) = 1 - \frac{1}{8\sqrt{2}}\,1\sqrt{1}\,(10 - 3\times 1) = 1 - \frac{7}{8\sqrt{2}} \simeq 0.381.$$

(d) Since 30 seconds is half a minute, you are asked for $P(1/2 \le X \le 1)$. This is

$$\begin{aligned}P(1/2 \le X \le 1) = F(1) - F(1/2) &= \frac{7}{8\sqrt{2}} - \frac{1}{8\sqrt{2}} \times \frac{1}{2}\sqrt{\frac{1}{2}}\left(10 - 3\times\frac{1}{2}\right) \\ &= \frac{1}{8\sqrt{2}}\left(7 - \frac{1}{2\sqrt{2}}\,\frac{17}{2}\right) = \frac{1}{8\sqrt{2}}\,\frac{1}{4\sqrt{2}}\left(28\sqrt{2} - 17\right) \\ &= \frac{1}{64}\left(28\sqrt{2} - 17\right) \simeq 0.353.\end{aligned}$$

Solutions to exercises

Solution to Exercise 1

(a) If the die is unbiased, all four outcomes are equally likely. So the probability that the die will come to rest on any particular face is 1/4. Thus the probability that it will come to rest on the face labelled 3 is 1/4.

(b) If all eight outcomes are equally likely, then the probability that the die will come to rest on any particular face is 1/8. So the probability that it will come to rest on a face labelled either 3 or 6 is 2/8 or 1/4.

(c) The two die rolls are independent, so the probability that the tetrahedral die comes to rest on the face labelled 3, and the octahedral die comes to rest on a face labelled 3 or 6, is the product of the probabilities from parts (a) and (b). So the required probability is $1/4 \times 1/4 = 1/16$.

Solution to Exercise 2

(a) An estimate of the probability that a tiger beetle found in the spring will be bright red is

$$\frac{302}{302+202} = \frac{302}{504} \simeq 0.60.$$

(b) An estimate of the probability that a tiger beetle found in the summer will not be bright red is

$$\frac{95}{72+95} = \frac{95}{167} \simeq 0.57.$$

(c) A direct estimate of the probability that a tiger beetle found in the summer will be bright red is

$$\frac{72}{72+95} = \frac{72}{167} \simeq 0.43.$$

Alternatively, using the probability rule for complementary events,

$$\begin{aligned} &P(\text{a tiger beetle found in the summer will be bright red}) \\ &\quad = 1 - P(\text{a tiger beetle found in the summer will not be bright red}) \\ &\quad = 1 - \frac{95}{167} = \frac{72}{167} \simeq 0.43. \end{aligned}$$

Solution to Exercise 3

(a) This is a count so it requires a discrete probability model.

(b) This is a measurement so it requires a continuous probability model.

(c) This is a measurement so it requires a continuous probability model.

(d) This is a count so it requires a discrete probability model.

Solution to Exercise 4

(a) The die is equally likely to come to rest on each of the four faces, so the probability that it lands on any particular face is 1/4. The probability function of X, the score on the face on which it comes to rest, is

$$p(x) = 1/4, \quad x = 1, 2, 3, 4.$$

(b) The probability that an octahedral die comes to rest on any particular face is 1/8, so the probability function of Y, the score on the face on which it comes to rest, is

$$p(y) = 1/8, \quad y = 1, 2, \ldots, 8.$$

Solution to Exercise 5

(a) The histogram suggests two main peaks, one at approximately 1.75 minutes, the other at around 4 minutes; it is bimodal.

(b) A possible curve is sketched in Figure 30.

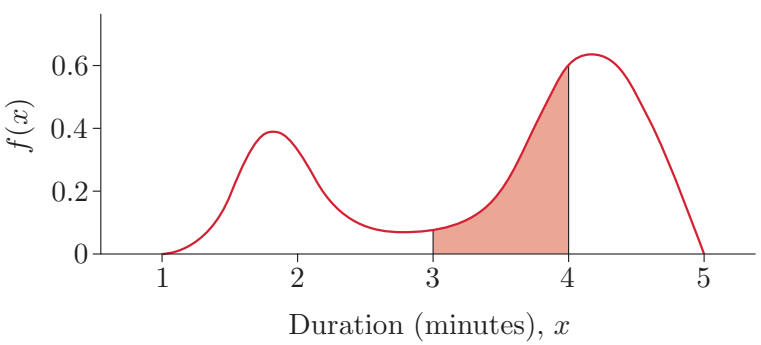


Figure 30 Possible model for durations of eruptions

If the total area under the curve is 1, then the shaded area represents the probability that an eruption will last between 3 and 4 minutes.

Solution to Exercise 6

(a) $$\int_{-1}^{2} x^3(1-x)\,dx = \int_{-1}^{2} \left(x^3 - x^4\right) dx = \left[\frac{x^4}{4} - \frac{x^5}{5}\right]_{-1}^{2}$$
$$= \frac{2^4}{4} - \frac{2^5}{5} - \left(\frac{(-1)^4}{4} - \frac{(-1)^5}{5}\right)$$
$$= 4 - \frac{32}{5} - \frac{1}{4} - \frac{1}{5} = -\frac{57}{20} = -2.85.$$

(b) $$\int_{1}^{2} \left(3x^2 - x^{-2}\right) dx = \left[\frac{3x^3}{3} - \frac{x^{-1}}{-1}\right]_{1}^{2} = \left[x^3 + \frac{1}{x}\right]_{1}^{2}$$
$$= 2^3 + \frac{1}{2} - \left(1^3 + \frac{1}{1}\right)$$
$$= 8 + \frac{1}{2} - 1 - 1 = \frac{13}{2} = 6.5.$$

Solution to Exercise 7

(a) $P\left(\frac{1}{4} \leq X \leq \frac{1}{2}\right) = \displaystyle\int_{\frac{1}{4}}^{\frac{1}{2}} (2-2x)\,dx = \Big[2x - x^2\Big]_{\frac{1}{4}}^{\frac{1}{2}}$

$$= 1 - \frac{1}{4} - \left(\frac{1}{2} - \frac{1}{16}\right) = \frac{5}{16}.$$

(b) $P(2 \leq X \leq 5) = \displaystyle\int_2^5 \frac{3}{125}(x-1)^2\,dx = \frac{3}{125}\int_2^5 (x^2 - 2x + 1)\,dx$

$$= \frac{3}{125}\left[\frac{x^3}{3} - x^2 + x\right]_2^5$$

$$= \frac{3}{125}\left\{\frac{125}{3} - 25 + 5 - \left(\frac{8}{3} - 4 + 2\right)\right\}$$

$$= \frac{3}{125}\left(\frac{117}{3} - 18\right) = \frac{63}{125} = 0.504.$$

Solution to Exercise 8

(a) 'P.d.f. 1' is not a valid p.d.f. because it is negative over part of the range of X (in fact, for all values of x smaller than $1/2$).

(b) 'P.d.f. 2' is non-negative over the range of X but is not a valid p.d.f. for any finite value of K. This is because

$$\int_0^1 \frac{1}{K}\,x^{-2}\,dx = \left[-\frac{1}{Kx}\right]_0^1 = \frac{1}{K}\{-1 - (-\infty)\} = \infty.$$

(c) 'P.d.f. 3' is non-negative over the range of X; this is because it is a linear function joining the values $f(0) = 3$ and $f(1) = 3/2$. It is not a valid p.d.f., however, because $\int_0^1 f(x)\,dx \neq 1$:

$$\int_0^1 \frac{3}{2}(2-x)\,dx = \frac{3}{2}\int_0^1 (2-x)\,dx = \frac{3}{2}\left[2x - \frac{x^2}{2}\right]_0^1$$

$$= \frac{3}{2}\left\{2 - \frac{1}{2} - (0-0)\right\} = \frac{3}{2} \times \frac{3}{2} = \frac{9}{4}.$$

(d) 'P.d.f. 4' is a valid p.d.f. It is non-negative over the range of X; this is because it is a linear function joining the values $f(0) = 4/3$ and $f(1) = 2/3$. Also,

$$\int_0^1 \frac{2}{3}(2-x)\,dx = \frac{2}{3}\int_0^1 (2-x)\,dx = \frac{2}{3} \times \frac{3}{2} = 1.$$

Here, we used $\int_0^1 (2-x)\,dx = \frac{3}{2}$ from the solution to part (c).

Solution to Exercise 9

(a) You found the probability mass function $p(y)$ for Y, the score on an octahedral die, in Exercise 4(b). This is shown in the following table, together with the cumulative distribution function $F(y)$. The c.d.f. was found by summing values of the p.m.f.

Table 23

y	1	2	3	4	5	6	7	8
$p(y)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
$F(y)$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{5}{8}$	$\frac{3}{4}$	$\frac{7}{8}$	1

(b) (i) $P(Y \leq 3) = F(3) = \frac{3}{8}$.

(ii) $P(Y < 6) = P(Y \leq 5) = F(5) = \frac{5}{8}$.

(iii) $P(Y > 4) = 1 - P(Y \leq 4) = 1 - F(4) = 1 - \frac{1}{2} = \frac{1}{2}$.

(iv) $P(Y \geq 4) = 1 - P(Y \leq 3) = 1 - F(3) = 1 - \frac{3}{8} = \frac{5}{8}$.

Solution to Exercise 10

(a) You are told that f is non-negative for all x in its range. Also, $\int_3^6 f(x)\,dx = 1$. To see this,

$$\begin{aligned}\int_3^6 \frac{1}{30}(10x - x^2 - 14)\,dx &= \frac{1}{30}\int_3^6 (10x - x^2 - 14)\,dx \\ &= \frac{1}{30}\left[5x^2 - \frac{x^3}{3} - 14x\right]_3^6 \\ &= \frac{1}{30}\{180 - 72 - 84 - (45 - 9 - 42)\} \\ &= \frac{1}{30}(24 + 6) = 1.\end{aligned}$$

Therefore f is a valid p.d.f.

(b) For $3 < x < 6$,

$$\begin{aligned}F(x) &= \frac{1}{30}\int_3^x (10y - y^2 - 14)\,dy = \frac{1}{30}\left[5y^2 - \frac{y^3}{3} - 14y\right]_3^x \\ &= \frac{1}{30}\left\{5x^2 - \frac{x^3}{3} - 14x - (45 - 9 - 42)\right\} \\ &= \frac{1}{30}\left(5x^2 - \frac{x^3}{3} - 14x + 6\right) = \frac{1}{90}\left(15x^2 - x^3 - 42x + 18\right).\end{aligned}$$

(c) You are asked for $P(X < 4)$. This is

$$P(X < 4) = F(4) = \frac{1}{90}(240 - 64 - 168 + 18) = \frac{26}{90} = \frac{13}{45} \simeq 0.289.$$

(d) You are asked for $P(4 \leq X \leq 5)$. This is

$$\begin{aligned}P(4 \leq X \leq 5) &= F(5) - F(4) \\ &= \frac{1}{90}(375 - 125 - 210 + 18) - \frac{26}{90} \\ &= \frac{1}{90}(58 - 26) = \frac{32}{90} = \frac{16}{45} \simeq 0.356.\end{aligned}$$

Here, we used $F(4) = 26/90$ from the solution to part (c).

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 75: © Yarygin / www.istockphoto.com

Page 77 top: © Öterreichische Nationalbibliothek

Page 77 bottom: © elen1 / www.123rf.com

Page 78 bottom: This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 79: © rido / www.123rf.com

Page 81: © Scott Rothstein / www.istockphoto.com

Page 82: © Whitney Cranshaw, Colorado State University, Bugwood.org

Page 83: Sigismund von Dobschtz This file is licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License https://creativecommons.org/licenses/by-sa/3.0/deed.en

Page 84: PhotosIndia.com LLC / www.123rf.com

Page 88: chirawan Somsanuk / www.123rf.com

Page 91: © rozbyshaka / www.istockphoto.com

Page 94: © stug.stug This file is licensed under the Creative Commons Attribution-ShareAlike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 97: Arno Kohlem / https://en.wikipedia.org/wiki/File:Perth, Kwinana_freeway.jpg This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 102: © iStockphoto.com / RolandBlunck

Page 109: Leonardo Patrizi / www.istockphoto.com

Page 112: Martin Damen / www.123rf.com

Page 115: Courtesy of Texas A&M University

Page 116: Taken from: http://www.pakwheels.com/forums/attachments/spotting-hobbies-other-stuff/469848d1140147840-can-u-fit-so-many-people-ur-car-mostinmini_mjc_pakwheels-com-.jpg

Page 122: Duncan Noakes / www.123rf.com

Page 123: © Gerdzhikov / www.istockphoto.com

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.